

Natural Induction: An Objective Bayesian Approach

James O. Berger, José M. Bernardo and Dongchu Sun

Abstract. The statistical analysis of a sample taken from a finite population is a classic problem for which no generally accepted objective Bayesian results seem to exist. Bayesian solutions to this problem may be very sensitive to the choice of the prior, and there is no consensus as to the appropriate prior to use.

This paper uses new developments in reference prior theory to justify and generalize Perks (1947) ([15]) ‘rule of succession’ — determining the probability that a new element from a population will have a property, given that *all* n previous elements from a random sample possessed the property — and to propose a new objective Bayesian solution to the ‘law of natural induction’ problem — determining the probability that all elements in a finite population have the property, given that all previous elements had the property.

The prior used for the first problem is the reference prior for an underlying hypergeometric probability model, a prior first suggested by Jeffreys (1946) ([10]) and recently justified on the basis of an exchangeability argument in Berger, Bernardo and Sun (2009) ([4]). The reference prior in the second problem arises as a modification to this prior that results from declaring the quantity of interest to be whether or not all the elements in the finite population have the property under scrutiny.

Inducción en las Ciencias de la Naturaleza: Una Solución Bayesiana Objetiva

Resumen. El análisis estadístico de una muestra aleatoria extraída de una población finita es un problema clásico para el que no parece existir una solución bayesiana generalmente aceptada. Las soluciones bayesianas a este problema pueden ser muy sensibles a la elección de la distribución inicial, y no existe consenso sobre la distribución inicial que debería ser utilizada.

En este trabajo se hace uso de desarrollos recientes del análisis de referencia para justificar y generalizar la solución de Perks (1947) ([15]) a la ‘regla de sucesión’ — la probabilidad de que un nuevo elemento de la población tenga una propiedad si *todos* los elementos de una muestra aleatoria la tienen — y para proponer una nueva solución bayesiana objetiva a la ‘ley de inducción natural’, — la probabilidad de que todos los elementos de una población finita tengan una propiedad si todos los elementos de la muestra la tienen. La distribución inicial utilizada para el primer problema es la distribución de referencia para el modelo probabilístico hipergeométrico subyacente, una distribución inicial sugerida por Jeffreys (1946) ([10]) y recientemente justificada utilizando un argumento de intercambiabilidad en Berger, Bernardo and Sun (2009) ([4]). La distribución de referencia para el segundo problema se obtiene

Presentado por / Submitted by Francisco Javier Girón González-Torre.

Recibido / Received: 12 de febrero de 2009. Aceptado / Accepted: 4 de marzo de 2009.

Palabras clave / Keywords: Bayes factors, Binomial model, Finite parameter space, Finite population sampling, Jeffreys prior, Hypergeometric model, Laplace law of succession, Natural induction, Objective Bayes, Precise hypothesis testing, Objective prior, Reference analysis, Reference prior.

Mathematics Subject Classifications: Primary: 62A01, 62F15; Secondary: 62B10, 62P10, 62P30, 62P35

© 2009 Real Academia de Ciencias, España.

como resultado de modificar la distribución anterior al declarar que el problema de interés es determinar si es o no es cierto que todos los elementos de una población finita tienen la propiedad objeto de estudio.

1 The Problem

The “rule of succession” and “law of natural induction” (see below for definitions) have been discussed for hundreds of years by scientists, philosophers, mathematicians and statisticians. Zabell (1989, 2005) ([18, 19]) gives introductions to much of the more quantitative side of this history, as well as providing numerous modern insights.

Our focus here is primarily technical: to produce the specific “rule” and “law” that arise from adopting the *reference prior* approach to objective Bayesian analysis, as this approach has proven itself to be quite successful in a wide variety of contexts (see Bernardo, 1979, 2005 ([5, 7]); Berger and Bernardo, 1992 ([2]); and Berger, Bernardo and Sun, 2009 ([3, 4]), for discussion).

Sampling from a finite population. The most common statistical framework in which these subjects are discussed is that of a finite population of size N , where the interest centers on R , the unknown number of elements from the population which share a certain property. For instance, the population may consist of a batch of N recently produced items, R of which satisfy the required specifications and may therefore be safely sold, or it may consist of a population of N individuals, R of which share some genetic characteristic. The elements which share the property under analysis will be called *conforming*, and the event that a particular element in the population is conforming will be denoted by $+$. Given the information provided by a random sample of size n (without replacement) from the population which has yielded r conforming items, interest usually centers in one of these problems:

- The proportion $\theta = R/N$ of conforming items in the population.
- The probability $\Pr(+ | r, n, N)$ that an element randomly selected among the remaining unobserved $N - n$ elements turns out to be conforming. The particular case $\Pr(+ | r = n, n, N)$, that the next observed item is conforming, given that the first n observed elements are conforming, is commonly referred to as the *rule of succession*.
- The probability $\Pr(\text{All } + | n, N)$ that *all* the elements in the population are conforming given that the first n observed elements are conforming. This is commonly referred to as the *law of natural induction*.

The probability model for the relevant sampling mechanism is clearly hypergeometric, so that

$$\Pr(r | n, R, N) = \text{Hy}(r | n, R, N) = \frac{\binom{R}{r} \binom{N-R}{n-r}}{\binom{N}{n}}, \quad (1)$$

where $r \in \{0, \dots, \min(R, n)\}$, $0 \leq n \leq N$, and $0 \leq R \leq N$.

Bayesian solutions to the problems described above require specification of a prior distribution $\Pr(R|N)$ over the unknown number R of conforming elements in the population. As will later become evident, these solutions are quite sensitive to the particular choice of the prior $\Pr(R|N)$. Note, however, that non-Bayesian solutions to these problems are very problematical, in part because of the discreteness of the problem and the fact that the interesting data outcome — all n observed elements are conforming — is extreme and, in part, because of structural difficulties with non-Bayesian approaches in *confirming* a precise hypothesis such as “all the elements in a population are conforming.”

Predictive and posterior distributions. If $\Pr(R|N)$, $R = 0, \dots, N$, defines a prior distribution for R , the posterior probability of R conforming elements in the population having observed r conforming elements within a random sample of size n is, by Bayes theorem,

$$\Pr(R | r, n, N) = \frac{\text{Hy}(r | n, R, N) \Pr(R | N)}{\Pr(r | n, N)}, \quad (2)$$

for $R \in \{r, \dots, N - n + r\}$, and zero otherwise, where

$$\Pr(r | n, N) = \sum_{R=r}^{N-n+r} \text{Hy}(r | n, R, N) \Pr(R | N) \quad (3)$$

is the predictive distribution for the number r of conforming elements in a random sample of size n . Since, given N , the discrete parameter $\theta = R/N$, for $\theta \in \{0, 1/N, \dots, 1\}$, is a one-to-one transformation of R , the corresponding posterior of θ is just $\pi(\theta | r, N, n) = \Pr(N\theta | r, N, n)$.

Rule of succession. By the total probability theorem, the probability that an element randomly selected among the remaining unobserved $N - n$ elements is conforming is

$$\Pr(+ | r, n, N) = \sum_{R=r}^{N-n+r} \frac{R - r}{N - n} \Pr(R | r, n, N). \quad (4)$$

In particular, the probability of the event E_n that something which has occurred n times and has not hitherto failed to occur (so that $r = n$) will occur again is

$$\Pr(E_n | N) = \Pr(+ | r = n, n, N) \quad (5)$$

which, for many commonly used priors, turns out to be independent of the population size N . This is commonly referred to as the rule of succession. As n increases, $\Pr(E_n | N)$ converges quickly to one for all N for commonly used priors. This agrees with the usual perception that, if an event has been observed for a relatively large uninterrupted number of times, it is very likely that it will be observed again in the next occasion.

Law of natural induction. The posterior probability that all the N elements in the population are conforming given that all the n elements in the sample are, is

$$\Pr(\text{All } + | n, N) = \Pr(R = N | r = n, n, N). \quad (6)$$

In typical applications, n will be moderate and N will be much larger than n . For many conventional priors, $\Pr(\text{All } + | n, N)$ would then be very small, and this clearly conflicts with the common perception from scientists that, as n increases, $\Pr(\text{All } + | n, N)$ should converge to one, whatever the (often very large) value of N might be. A formal objective Bayesian solution to this problem, typically known as the law of natural induction, is the main objective of this paper, and will require following a suggestion of Jeffreys.

2 Conventional Objective Bayesian Inference

Both Bayes (1763) [1] and Laplace (1774, 1812) [13, 14] utilized a constant prior for unknowns. Since R may take $N + 1$ different values, the constant prior is thus the uniform distribution

$$\pi_u(R | N) = \frac{1}{N + 1}, \quad R = 0, \dots, N. \quad (7)$$

We first review the analysis for this conventional objective prior, as given in Broad (1918) ([9]).

Predictive and posterior distributions. The corresponding predictive distribution $\pi_u(r | n, N)$ for the number r of conforming elements in a random sample of size n is

$$\pi_u(r | n, N) = \frac{1}{N + 1} \sum_{R=r}^{N-n+r} \text{Hy}(r | n, R, N) = \frac{1}{n + 1},$$

for $r = 0, \dots, n$, a uniform distribution over its $n + 1$ possible values, which is therefore independent of N . Substituting into (2), the corresponding posterior distribution for R is

$$\pi_u(R | r, n, N) = \frac{n + 1}{N + 1} \text{Hy}(r | n, R, N) = \frac{\binom{R}{r} \binom{N-R}{n-r}}{\binom{N+1}{n+1}}, \quad (8)$$

for $R \in \{r, \dots, N - n + r\}$, and zero otherwise. In particular, the posterior probability that all the N elements in the population are conforming, given that all the n elements in the sample are conforming, is

$$\pi_u(\text{All} + | n, N) = \pi_u(R = N | r = n, n, N) = \frac{n + 1}{N + 1}, \quad (9)$$

which is essentially the ratio of the sample size to the population size. Notice that, when n is much smaller than N as will often be the case, $\pi_u(\text{All} + | n, N)$ will be close to zero even for large values of the sample size n .

Law of succession. From (4), the probability that an element randomly selected from among the remaining unobserved $N - n$ elements is conforming given the uniform prior (7), so that $\Pr(R | r, n, N)$ is given in (8), reduces to

$$\pi_u(+ | r, n, N) = \sum_{R=r}^{N-n+r} \frac{R - r}{N - n} \frac{\binom{R}{r} \binom{N-R}{n-r}}{\binom{N+1}{n+1}} = \frac{r + 1}{n + 2}. \quad (10)$$

This is usually known as Laplace's rule of succession, although Laplace (1774) ([13]) did not consider the case of finite N and only derived (10) for its continuous binomial approximation without apparently realizing that this is also an exact expression for finite N , a result established by Broad (1918) ([9]). In particular, with the uniform prior (7), the probability of the event E_n that something which has occurred n times and has not hitherto failed to occur (so that $r = n$) will occur again is

$$\pi_u(E_n) = \pi_u(+ | r = n, n, N) = \frac{n + 1}{n + 2}, \quad (11)$$

which is independent of the population size N . As n increases, $\pi_u(E_n)$ quickly converges to one.

Notice the dramatically different behaviour of the seemingly related Equations (9) and (11). In typical applications, n will be moderate and N will be much larger than n ; if this is the case, $\pi_u(\text{All} + | n, N)$ will be close to zero, but $\pi_u(E_n)$ will be close to one. Thus, if an event has been observed for a relatively large number of uninterrupted times, and the uniform prior (7) is used for both problems, one obtains that it is very likely that it will be observed again in the next occasion, but quite unlikely that it will *always* be observed in the future.

3 Reference Analysis of the Hypergeometric Model

The conventional use of a uniform prior for discrete parameters can ignore the structure of the problem under consideration. For the hypergeometric model, where the values of R are actual numbers, not merely labels, there is arguably a clear structure. Indeed, for large N it is well known that the hypergeometric distribution is essentially equivalent to the binomial distribution with parameter $\theta = R/N$. The objective prior for R should thus be compatible with the appropriate objective prior for θ in a Binomial $\text{Bi}(r | n, \theta)$ model; this is commonly chosen to be the corresponding Jeffreys (and reference) prior, which is the (proper) distribution

$$\pi_r(\theta) = \text{Be}(\theta | \frac{1}{2}, \frac{1}{2}) = \frac{1}{\pi} \frac{1}{\sqrt{\theta(1-\theta)}}, \quad 0 < \theta < 1. \quad (12)$$

As N increases, the uniform prior (7) remains uniform, and this is clearly not compatible with (12).

Similar reasoning led Jeffreys (1946, 1961) ([10, 12]) to suppose that the R conforming items arose as a random sample from a binomial population with parameter p (which can be thought of as the limiting value of $\theta = R/N$ as $N \rightarrow \infty$), and then assign p the Jeffreys prior in (12). This hierarchical structure can be justified on the basis of exchangeability, as observed in Berger, Bernardo and Sun (2009) ([4]) (which also generalized the approach to other discrete distributions). The resulting induced reference prior for R is

$$\pi_r(R|N) = \frac{1}{\pi} \frac{\Gamma(R + \frac{1}{2}) \Gamma(N - R + \frac{1}{2})}{\Gamma(R + 1) \Gamma(N - R + 1)}, \quad R \in \{0, 1, \dots, N\}, \quad (13)$$

which may also be written as

$$\pi_r(R|N) = f(R) f(N - R), \quad R \in \{0, 1, \dots, N\}, \quad (14)$$

where

$$f(y) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(y + 1/2)}{\Gamma(y + 1)}, \quad y \geq 0. \quad (15)$$

As will later become evident, the positive, strictly decreasing function defined by (15) occurs very frequently in the derivations associated to the problems analyzed in this paper. Since $\Gamma(1/2) = \sqrt{\pi}$, $f(0) = 1$. Moreover, using Stirling's approximation to the Gamma functions, it is easily seen that, for large y ,

$$f(y) \approx \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{y}},$$

so that if R and $N - R$ are both large, one has

$$\pi_r(R|N) \approx \frac{1}{\pi} \frac{1}{\sqrt{R(N - R)}}.$$

For $N = 1$, the reference prior (14) is the uniform prior $\pi_r(R|N = 1) = \{1/2, 1/2\}$, for $R \in \{0, 1\}$, as one would certainly expect, but this is the *only* case where the reference prior for the hypergeometric is uniform. Indeed, using Stirling's approximation for the Gamma functions in (13) one gets, in terms of $\theta = R/N$,

$$\pi_r(\theta|N) \approx \frac{1}{N + \frac{2}{\pi}} \text{Be}\left(\frac{N\theta + \frac{1}{\pi}}{N + \frac{2}{\pi}} \mid \frac{1}{2}, \frac{1}{2}\right), \quad \theta = 0, 1/N, \dots, 1, \quad (16)$$

which is basically proportional to $\text{Be}(\theta | \frac{1}{2}, \frac{1}{2})$, and hence compatible with the reference prior for the continuous limiting model $\text{Bi}(r|n, p)$ with $p = \lim_{N \rightarrow \infty} R/N$.

Reference predictive and posterior distributions. Using (3), the reference prior predictive distribution of the number r of conforming items in a random sample of size n is

$$\begin{aligned} \Pr(r|n, N) &= \sum_{R=0}^N \text{Hy}(r|R, N, n) \pi_r(R|N) \\ &= \frac{1}{\pi} \frac{\Gamma(r + \frac{1}{2}) \Gamma(n - r + \frac{1}{2})}{\Gamma(r + 1) \Gamma(n - r + 1)} \\ &= f(r) f(n - r) = \pi_r(r|n), \end{aligned}$$

which is independent of the population size N . Notice that, as in the case of the uniform prior, the reference prior predictive distribution of r given n has precisely the same mathematical form as the reference prior of R given N , $\pi_r(R|N)$.

Furthermore, using (2) and the last result, the reference posterior distribution of R turns out to be

$$\begin{aligned} \pi_r(R|r, n, N) &= \frac{c(r, n, N) \Gamma(R + \frac{1}{2}) \Gamma(N - R + \frac{1}{2})}{\Gamma(R - r + 1) \Gamma(N - R - (n - r) + 1)}, \quad (17) \\ c(r, n, N) &= \frac{\Gamma(n + 1) \Gamma(N - n + 1)}{\Gamma(N + 1) \Gamma(r + \frac{1}{2}) \Gamma(n - r + \frac{1}{2})}. \end{aligned}$$

In particular, if $N = 2$ and $n = 1$, this yields

$$\pi_r(R | r = 0, n = 1, N = 2) = \{3/4, 1/4, 0\}, \quad R = 0, 1, 2,$$

which may be compared with the corresponding result $\{2/3, 1/3, 0\}$, obtained from a uniform prior.

Substituting $R = N$ and $r = n$ into (17) and simplifying, the reference posterior probability that all elements in the population are conforming, given that all elements in the sample are conforming, is

$$\pi_r(\text{All} + | n, N) = \frac{\Gamma(N + 1/2)}{\Gamma(N + 1)} \frac{\Gamma(n + 1)}{\Gamma(n + 1/2)} = \frac{f(N)}{f(n)} \approx \sqrt{\frac{n}{N}}. \quad (18)$$

Thus, $\pi_r(\text{All} + | n, N)$ is basically the square root of the ratio of the sample size to the population size, a considerable contrast to the result $(n + 1)/(N + 1)$ obtained in (9) for the uniform prior.

Rule of succession. From (4), the probability that an element randomly selected from among the remaining unobserved $N - n$ elements is conforming, for the reference prior (13), reduces to

$$\pi_r(+ | r, n, N) = \sum_{R=r}^{N-n+r} \frac{R-r}{N-n} \pi_r(R | r, n, N) = \frac{r + 1/2}{n + 1}, \quad (19)$$

which is independent of N . Equation (19) provides the reference rule of succession, which was first obtained by Perks (1947) ([15]) although, following Laplace, he only derived it for the limiting binomial approximation ($N = \infty$ case), apparently not realizing that it was also an exact expression for any finite population size N .

The reference rule of succession (19) may be compared with Laplace's $(r + 1)/(n + 2)$ of (10). In particular, the corresponding reference probability of the event E_n — that something which has occurred n times and has not hitherto failed to occur will occur — is, for any population size N ,

$$\pi_r(E_n) = \pi_r(+ | r = n, n, N) = \frac{n + 1/2}{n + 1}. \quad (20)$$

As one would require, for $n = 0$ (and hence with no initial information), both (11) and (20) yield $1/2$. For $n = 1$, Laplace yields $2/3$ while the corresponding reference probability is $3/4$. It is easily verified that, as n increases, the reference law of succession (20) has an appreciably faster convergence to one than Laplace's (11).

4 Natural Induction

In line with his early discussions on scientific enquiry (Wrinch and Jeffreys, 1921-23 ([17]), later expanded in Jeffreys, 1931 ([11])), Jeffreys (1961, p. 128) ([12]) disagreed with the result (9) — and would also have disagreed with (18) — arguing that, to justify natural induction, one should be able to demonstrate that a law is probably correct for all elements of a population of size N , given that it has proven to be correct in all of a very large number n of randomly chosen instances, even if n is appreciably smaller than N . In contrast, (9) and (18) can be quite small for large n if N is much larger than n . (Note that both (11) and (20) are near 1 for large n , but these probabilities refer to the event E_n that a further randomly chosen element will obey the stated law, not to the event that *all* elements in the population obey that law.)

To correct this problem, Jeffreys argued that the prior probability that all elements of the population have the property, $\Pr(R = N)$, must be some fixed value independent of N . He argued that this is reasonable, asserting that any clearly stated natural law has a positive prior probability of being correct, and he made several specific proposals (Jeffreys, 1961, Sec. 3.2 ([12])) for the choice of prior probability. The simplest choice is to let $\Pr(R = N) = 1/2$, and this is the choice arising from the reference analysis below. For a recent review of Jeffreys' ideas, see Robert, Chopin and Rousseau (2009) ([16]).

Reference analysis. A solution to the natural induction problem that satisfies the scientific desiderata described by Jeffreys may be obtained from a standard use of reference analysis, if the parameter of interest is chosen to be whether or not $R = N$, rather than the actual value of R . The result, described below, can also be phrased in terms of testing the hypothesis that $R = N$ versus the alternative $R \neq N$.

Lemma 1 Define the parameter of interest (in the reference prior analysis) to be

$$\phi = \begin{cases} \phi_0 & \text{if } R = N \text{ (All +)}, \\ \phi_1 & \text{if } 0 \leq R < N. \end{cases}$$

Then the corresponding reference prior, $\pi_\phi(R | N)$, of the unknown parameter R is

$$\pi_\phi(R | N) = \begin{cases} \frac{1}{2} & \text{if } R = N \\ \frac{1}{2} \frac{f(R) f(N-R)}{1-f(N)} & \text{if } 0 \leq R < N, \end{cases}$$

where $f(y)$ is defined in (15).

PROOF. To have a representation of the unknown parameter R in terms of the quantity of interest ϕ and a nuisance parameter λ , define

$$\lambda = \begin{cases} \lambda_0 & \text{if } R = N \text{ (All +)}, \\ R & \text{if } 0 \leq R < N. \end{cases}$$

The sampling distribution of the data r given $\phi = \phi_1$ is still the density

$$\Pr(r | \phi = \phi_1, \lambda \neq \lambda_0, n, N) = \text{Hy}(r | n, R, N),$$

but now R is restricted to the set $\{0, 1, \dots, N - 1\}$. The reference prior of a model with a restricted parameter space is obtained as the restriction of the reference prior from the unrestricted model. Hence, using the fact that

$$\pi_r(N | N) = \frac{\Gamma(N + 1/2)}{\sqrt{\pi} \Gamma(N + 1)} = f(N),$$

the conditional reference prior of λ given $\phi = \phi_1$ is the renormalized version of (14)

$$\pi(\lambda | \phi = \phi_1, N) = \frac{f(R) f(N - R)}{1 - f(N)}. \quad (21)$$

On the other hand, $\pi(\lambda = \lambda_0 | \phi = \phi_0, N) = 1$ since, given $\phi = \phi_0$, the nuisance parameter λ must be equal to λ_0 . Moreover, ϕ has only two possible values and, therefore, its marginal reference prior is simply $\pi(\phi = \phi_0) = \pi(\phi = \phi_1) = 1/2$. Hence, the joint reference prior of the unknown parameter R , when ϕ is the quantity of interest, is $\pi_\phi(R | N) = \pi(\lambda | \phi, N) \pi(\phi)$, which yields the conclusion. ■

Using this prior, the required reference posterior probability of the event $\{\text{All +}\}$ that all N elements in the population are conforming ($R = N$), given that all r elements in a random sample of size n are conforming ($r = n$), is found in the following result.

Theorem 1

$$\pi_\phi(\text{All +} | n, N) = \left(1 + \frac{f(n) - f(N)}{1 - f(N)} \right)^{-1}, \quad (22)$$

where $f(y)$ is defined in (15).

PROOF. Note that

$$\begin{aligned}\pi_\phi(\text{All} + |n, N) &= \pi_\phi(\phi = \phi_0 | r = n, N) \\ &= \frac{\frac{1}{2} \Pr(r = n | \phi = \phi_0, n, N)}{\frac{1}{2} \Pr(r = n | \phi = \phi_0, n, N) + \frac{1}{2} \Pr(r = n | \phi = \phi_1, n, N)} \\ &= (1 + \Pr(r = n | \phi = \phi_1, n, N))^{-1},\end{aligned}$$

since $\Pr(r = n | \phi = \phi_0, n, N)$ is obviously one. By the total probability theorem,

$$\Pr(n | \phi = \phi_1, n, N) = \sum_{R=n}^{N-1} \Pr(n | \phi = \phi_1, R, n, N) \pi_\phi(R | \phi = \phi_1, N)$$

and, using the fact that, for all N ,

$$\Pr(r = n | n, N) = \sum_{R=n}^N \text{Hy}(n | n, R, N) \pi_r(R | N, \mathcal{M}) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(n + \frac{1}{2})}{\Gamma(n + 1)} = f(n),$$

it is easily shown that

$$\Pr(r = n | R < N, n, N) = \frac{f(n) - f(N)}{1 - f(N)}.$$

The conclusion is immediate. ■

Hence, for large population sizes, $f(N) \approx 0$ and, as a consequence, the reference posterior probability, $\pi_\phi(R = N | r = n, N)$, that all N elements in the population are conforming, given that all elements in a sample of size n are conforming, is then essentially independent of N and given by

$$\pi_\phi(\text{All} + |n, N) \approx (1 + f(n))^{-1} = \left(1 + \frac{1}{\sqrt{\pi}} \frac{\Gamma(n + \frac{1}{2})}{\Gamma(n + 1)}\right)^{-1}, \quad (23)$$

which, for moderately large n , may further be approximated by

$$\pi_\phi(\text{All} + |n, N) \approx \frac{\sqrt{n}}{\pi^{-1/2} + \sqrt{n}}. \quad (24)$$

For instance, with $n = 100$ and $N = 1000$ the exact value of the required probability, given by (22), is $\pi_\phi(R = N | r = n, N) = 0.9623$, and the two approximations (23) and (24), respectively, yield 0.9467 and 0.9466.

Equation (22) may be compared with the result which one would obtain if a conventional uniform conditional prior

$$\pi_1(\lambda | \phi = \phi_1, N) = \pi_1(R | N - 1) = \frac{1}{N},$$

had been used instead of the structured conditional reference prior (21). It may be shown (Bernardo, 1985 ([6]), Bernardo and Smith, 1994, p. 322 ([8])) that this yields

$$\pi_1(\text{All} + |n, N) = \left(1 + \frac{1}{n+1} \left(1 - \frac{n}{N}\right)\right)^{-1}, \quad (25)$$

which is always larger than (22). For instance, with $n = 50$ and $N = 100$, the result in Theorem 1 yields $\pi_\phi(\text{All} + |n = 50, N = 100) = 0.9263$, while the use of (25) yields $\pi_1(\text{All} + |n = 50, N = 100) = 0.9903$, a much larger value. Thus the reference probability is considerably more conservative.

$N = \infty$ and hypothesis testing. As mentioned earlier, as $N \rightarrow \infty$, the hypergeometric $\text{Hy}(r | n, R, N)$ model converges to the binomial $\text{Bi}(r | n, p)$ model, with $p = \lim_{N \rightarrow \infty} R/N$. In this infinite population

setting, the model that all elements in the population have the specified property can be stated in the language of hypothesis testing as $H_0 \equiv \{p = 1\}$. The natural induction problem is thus formally that of finding the posterior probability of H_0 when all n observations have the property (i.e., are successes in the binomial model).

The reference prior assigns probability 1/2 to both $H_0 \equiv \{p = 1\}$ and $H_1 \equiv \{p < 1\}$ and, conditional on H_1 , assigns p the $\text{Be}(p | \frac{1}{2}, \frac{1}{2})$ reference prior. A straightforward computation then yields that the posterior probability of H_0 , given that the first n observations all have the specified property, is equal to

$$\Pr(p = 1 | n) = \left(1 + \frac{1}{\sqrt{\pi}} \frac{\Gamma(n + \frac{1}{2})}{\Gamma(n + 1)}\right)^{-1} = (1 + f(n))^{-1} \quad (26)$$

which, curiously, is the same expression as the approximation to the reference posterior probability $\pi_\phi(\text{All} + | n, N)$ obtained for large N in (23).

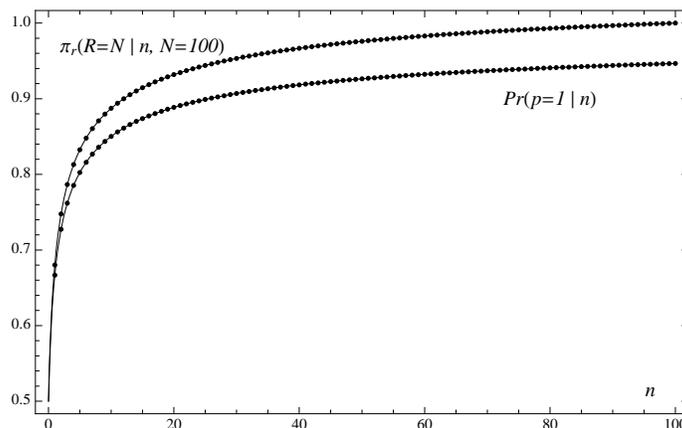


Figure 1. Reference posterior probabilities $\pi_\phi(\text{All} + | n, N = 100)$ that all the elements from a finite population of size $N = 100$ are conforming, given that all the elements from a random sample of size n are conforming, for $n = 0, 1, \dots, 100$, (above) and its Bayes factor continuous ($N = \infty$) approximation (below).

Figure 1 shows the exact values of (22) for $N = 100$ and $n = 0, 1, \dots, 100$, along with the binomial limiting case. Note that it is important to take the the population size, N , into account, except for very large population sizes. For example, if $r = 50$ and $N = 100$, the exact value of the required reference posterior probability that all elements in the population are conforming is $\pi_\phi(\text{All} + | n, N) = 0.9617$, but the continuous approximation gives only $\Pr(p = 1 | n) = 0.9263$.

5 Conclusions and an Example

The proposed solutions, based on reference prior theory, to the two originally posed problems can be summarized as follows.

The Reference Rule of Succession. In a population of N elements, from which n have been randomly sampled and been found to be conforming, the reference posterior probability that a new element, randomly selected from among the remaining unobserved $N - n$ elements, turns out to be conforming is

$$\pi_r(+ | r = n, n, N) = \frac{n + 1/2}{n + 1}, \quad (27)$$

which is independent of N and may be compared with Laplace's $(n + 1)/(n + 2)$.

The Reference Law of Natural Induction. If it is assumed that the parameter of interest is whether or not *all* elements of a population are conforming (i.e. that the property is a law of nature), then the reference posterior probability of this law is

$$\pi_{\phi}(\text{All} + | n, N) = \left(1 + \frac{\frac{1}{\sqrt{\pi}} \frac{\Gamma(n+1/2)}{\Gamma(n+1)} - \frac{1}{\sqrt{\pi}} \frac{\Gamma(N+1/2)}{\Gamma(N+1)}}{1 - \frac{1}{\sqrt{\pi}} \frac{\Gamma(N+1/2)}{\Gamma(N+1)}} \right)^{-1}. \quad (28)$$

For very large N (or infinite populations), this is essentially identical to (26).

Example. Many years ago, when visiting the Charles Darwin research station of the Galápagos islands, one of us (Bernardo) had a question posed by a zoologist. The zoologist had observed, and marked, 55 galápagos (tortoises) in one small island, all of which presented a particular shell modification, possibly related to their adaptation to the island vegetation. The zoologist asked for the probability that all galápagos in the island had that modification. He added that his sample was roughly random, and that he would estimate the total galápagos population in that island to be between 150 and 250 individuals.

At the time, Bernardo quoted the solution (25) based on a conditional uniform prior, which yielded the range [0.986, 0.989] (corresponding to the range of N). The reference probabilities (28) give the smaller values

$$\pi_{\phi}(\text{All} + | n = 55, N \in [150, 250]) \in [0.960, 0.970].$$

Note that these numbers are, nevertheless, appreciably higher than the limiting binomial approximation $\Pr(p = 1 | n = 55) = 0.929$, showing the importance of incorporating N into the analysis.

Finally, using (27), the zoologist could have been told that, if he went back to the island, he would have a reference posterior probability

$$\pi_r(+ | r = n, n = 55, N) = 0.991$$

that the first found unmarked galápagos also presented a modified shell.

Acknowledgement. J. O. Berger was supported by grant DMS-0635449 of the National Science Foundation, USA, J. M. Bernardo was supported by grant MTM2006-07801 of the Ministerio de Educación y Ciencia, Spain, and D. Sun was supported by grant SES-0720229 of the National Science Foundation, USA, and grants R01-MH071418 and R01-CA109675 of the National Institute of Health, USA.

References

- [1] BAYES, T., (1763). An essay towards solving a problem in the doctrine of chances. Published posthumously in *Phil. Trans. Roy. Soc. London*, **53**, 370–418 and **54**, 296–325.
- [2] BERGER, J. O. AND BERNARDO, J. M. (1992). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), Oxford: University Press, 35–60 (with discussion).
- [3] BERGER, J., BERNARDO, J. M. AND SUN, D. (2009). The formal definition of reference priors. *Ann. Statist.*, **37**, 905–938.
- [4] BERGER, J., BERNARDO, J. M. AND SUN, D. (2009). Reference priors for discrete parameter spaces. *Submitted*.
- [5] BERNARDO, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B*, **41**, 113–147 (with discussion). Reprinted in *Bayesian Inference*, **1** (G. C. Tiao and N. G. Polson, eds.) Oxford: Edward Elgar, 229–263.
- [6] BERNARDO, J. M. (1985). On a famous problem of induction. *Trab. Estadist.*, **36**, 24–30.

- [7] BERNARDO, J. M. (2005). Reference analysis. *Handbook of Statistics*, **25**, (D. K. Dey and C. R. Rao, eds.) Amsterdam: Elsevier, 17–90.
- [8] BERNARDO, J. M. AND SMITH, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.
- [9] BROAD, C. D. (1918). On the relation between induction and probability. *Mind*, **27**, 389–404.
- [10] JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Royal Soc. London A*, **186**, 453–461.
- [11] JEFFREYS, H. (1931). *Scientific Inference*. Cambridge: University Press.
- [12] JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford: University Press.
- [13] LAPLACE, P. S. (1774). Mémoire sur la probabilité des causes par les événements. *Mém. Acad. Sci. Paris*, **6**, 621–656. Reprinted in *Oeuvres Complètes* **8**, 27–68. Paris: Gauthier-Villars, 1891.
- [14] LAPLACE, P. S. (1812). *Théorie Analytique des Probabilités*. Paris: Courcier. Reprinted in *Oeuvres Complètes de Laplace*, **7**, 1878–1912. Paris: Gauthier-Villars.
- [15] PERKS, W. (1947). Some observations on inverse probability including a new indifference rule. *J. Inst. Actuaries* **73**, 285–334 (with discussion).
- [16] ROBERT, C. P., CHOPIN N. AND ROUSSEAU, J. (2009). Harold Jeffreys’ Theory of Probability revisited. *Statist. Science*, **23**, (to appear, with discussion).
- [17] WRINCH, D. AND JEFFREYS, H. (1921-23). On certain fundamental principles of scientific inquiry. *Philosophical Magazine* **42**, 369–390 and **45**, 368–374.
- [18] ZABELL, S. L. (1989). The rule of succession. *Erkenntnis*, **31**, 283–321.
- [19] ZABELL, S. L. (2005). *Symmetry and Its Discontents*. Cambridge: University Press.

James O. Berger

Dept. Statistical Science,
Duke University,
USA
berger@stat.duke.edu

José M. Bernardo

Dept. Estadística e I.O.,
Universitat de València,
Spain
jose.m.bernardo@uv.es

Dongchu Sun

Dept. Statistics,
University of Missouri-Columbia,
USA
sund@missouri.edu

Comments on: Natural Induction: An Objective Bayesian Approach

F. Javier Girón and Elías Moreno

We would like to thank the editor of RACSAM, Professor Manuel López Pellicer, for the opportunity he is offering to us of discussing this paper, and to congratulate Berger, Bernardo and Sun for an interesting and thought provoking paper.

The paper is motivated by the observation that the uniform prior for R , say $\pi(R|N) = 1/(N + 1)$, $R = 0, \dots, N$, gives poor results. It is shown that the posterior probability that all the N elements of the population are conforming, conditional on the event that all the observed n elements in the sample are conforming, is very small for N large, whatever moderate the sample size n should be. Then, a more reasonable prior $\pi(R|M)$ is provided on the ground of being compatible with the Jeffreys prior for the parameter θ of the Binomial limiting distribution with parameters (n, θ) , where $\theta = \lim_{R \rightarrow \infty, N \rightarrow \infty} R/N$. We enjoyed reading this clear argumentation.

However, in the abstract it is recognized that “*Bayesian solutions to this problem may be very sensitive to the choice of the prior, and there is no consensus as to the appropriate prior to use.*” It seems to us that the natural consequence of this assertion —that we share— is to consider a class of priors and reporting their posterior answers, instead of considering the posterior answer for the single reference prior for R . In this discussion we try to add the robustness analysis that we feel is missing in the paper.

For simplicity we will consider the limiting Binomial distribution $\text{Bi}(r|n, \theta)$, and the two problems addressed in the paper. Firstly, the testing problem

$$H_0 : \theta = 1 \text{ versus } H_1 : \theta \in [0, 1],$$

conditional on the dataset $r = n$, the event that all the elements of the sample are +. Secondly, the computation of the posterior predictive probability that a new observation is +, conditional on $r = n$.

The naive objective model selection formulation of this testing problem is that of choosing between the reduced sampling model

$$M_0 : \text{Bi}(n|n, \theta = 1)$$

and the full sampling model with the Jeffreys prior for θ , that is

$$M_1 : \left\{ \text{Bi}(n|n, \theta), \pi^J(\theta) = \frac{1}{\pi} \theta^{-1/2} (1 - \theta)^{-1/2} \right\}.$$

However, the Jeffreys prior does not concentrate its probability mass around the null with the consequence that those θ close to zero are privileged by the Jeffreys priors when being compared with the null $\theta = 1$. This is not reasonable, and many authors claim for a different prior to be used for testing that

Recibido / Received: 13 de marzo de 2009.

These comments refer to the paper of James O. Berger, José M. Bernardo and Dongchu Sun, (2009). **Natural Induction: An Objective Bayesian Approach**, *Rev. R. Acad. Cien. Serie A. Mat.*, 103(1), 125–135.

© 2009 Real Academia de Ciencias, España.

should be concentrated around the null. See, for instance, Jeffreys (1961, Chapter 5) ([8]), Gûnel and Dickey (1974) ([9]), who note that this is the ‘‘Savage continuity condition’’, Berger and Sellke (1987) ([3]), Casella and Berger (1987) ([4]), Morris (1987) ([12]), Berger (1994) ([2]), Casella and Moreno (2009) ([5]).

The point is how to define an objective class of priors that concentrate mass around the null. Fortunately, an answer to this question is provided by the class of intrinsic priors (Berger and Pericchi 1996 ([1]), Moreno *et al.* 1998 ([10])). This objective class of priors has been proved to behave extremely well for model selection in different contexts (Casella and Moreno 2006 ([5]), Consonni and La Roca 2008 ([7]), Moreno and Girón 2008 ([11])). The intrinsic priors for θ depend on a hyperparameter m that controls the degree of concentration of the priors around the null, and it ranges from 1 to n , so as to not exceed the concentration of the likelihood of θ (Casella and Moreno 2009 ([5])). For the above model selection problem standard calculations render the intrinsic prior class as the set of beta distributions $\text{Be}(m + 1/2, 1/2)$, that is

$$\pi^I(\theta|m) = \frac{\Gamma(m + 1)}{\Gamma(m + 1/2)\Gamma(1/2)}\theta^{m-1/2}(1 - \theta)^{-1/2}, \quad m = 1, 2, \dots, n.$$

Therefore, in the above model selection problem the Jeffreys prior should be replaced with the intrinsic prior, and M_0 should be compared with

$$M_1 : \{\text{Bi}(n|n, \theta), \pi^I(\theta|m), m = 1, 2, \dots, n\}.$$

We note that as m increases the intrinsic prior concentrates more around the null. Certainly, when the null is compared with models located in a small neighborhood of the null, one expects from the model selection problem an answer with more uncertainty than when the null is compared with models located far from it.

The posterior probability of the null for the intrinsic priors is given by

$$\Pr(\text{All} + |n, m) = \left(1 + \frac{\Gamma(m + 1)\Gamma(n + m + 1/2)}{\Gamma(m + 1/2)\Gamma(n + m + 1)}\right)^{-1}, \quad m = 1, \dots, n.$$

Likewise, the posterior probability that a new observation is +, conditional on $r = n$, is given by the total probability theorem as

$$\Pr(+|n, m) = \sum_{i=0}^1 \Pr(+|M_i, n, m)P(M_i|n, m),$$

where $\Pr(+|M_0, n, m) = 1$, and

$$\Pr(+|M_1, n, m) = \frac{n + m + 1/2}{n + m + 1}.$$

Example 1 *Assuming that the galápagos population in the island is large enough, we obtain that*

$$\min_{m=1, \dots, 55} \Pr(\text{All} + |n = 55, m) = \Pr(\text{All} + |n = 55, m = 55) = 0.586,$$

and

$$\max_{m=1, \dots, 55} \Pr(\text{All} + |n = 55, m) = \Pr(\text{All} + |n = 55, m = 1) = 0.869,$$

while

$$\Pr(+|n = 55, m) \simeq 0.998$$

for $m = 1, 2, \dots, 55$.

This example illustrates something about robustness that is well known: the posterior probability of an event is typically much less sensitive to the prior than the tests are. The posterior probability that a new observation is +, conditional on $r = n$, that we have obtained is similar to that given in the paper, but the report for the testing problem given in the paper and that given by us are rather different.

References

- [1] BERGER, J. O. AND PERICCHI, L. R., (1996). The intrinsic Bayes factor for model selection and prediction, *Journal of the American Statistical Association*, **91**, 109–122.
- [2] BERGER, J. O., (1994). An overview of robust Bayesian analysis (with discussion), *Test*, **3**, 5–124.
- [3] BERGER, J. O. AND SELKE, T. (1987). Testing a point null hypothesis: The irreconcilability of p-values and evidence (with discussion), *Journal of the American Statistical Association*, **82**, 112–122.
- [4] CASELLA, G. AND BERGER, R. L., (1987). Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem (with discussion), *Journal of the American Statistical Association*, **82**, 106–111.
- [5] CASELLA, G. AND MORENO E., (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association*, **101**, 157–167.
- [6] CASELLA, G. AND MORENO E., (2009). Assessing robustness of intrinsic test of independence in two-way contingency tables, *Journal of the American Statistical Association* (to appear).
- [7] CONSONNI, G. AND LA ROCA, L., (2008). Tests Based on Intrinsic Priors for the Equality of Two Correlated Proportions, *Journal of the American Statistical Association*, **103**, 1260–1269.
- [8] JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford: University Press.
- [9] GÛNEL, E. AND DICKEY, J., (1974). Bayes factors for independence in contingency tables, *Biometrika*, **61**, 545–557.
- [10] MORENO, E., BERTOLINO, F. AND RACUGNO, W. (1998). An Intrinsic Limiting Procedure for Model Selection and Hypotheses Testing, *Journal of the American Statistical Association*, **93**, 1451–1460.
- [11] MORENO, E. AND GIRÓN, F. J., (2008). Comparison of Bayesian objective procedure for variable selection in linear regression, *Test*, **17**, 472–492.
- [12] MORRIS, C. N., (1987). Discussion of Berger/Sellke and Casella/Berger. *Journal of the American Statistical Association*, **82**, 106–111.

F. Javier Girón

Departamento de Estadística e Investigación Operativa,
Universidad de Málaga,
España
fj_giron@uma.es

Elías Moreno

Departamento de Estadística e Investigación Operativa
Universidad de Granada,
España
emoreno@ugr.es

Comments on: Natural Induction: An Objective Bayesian Approach

Dennis V. Lindley

This very fine paper is valuable because it produces challenging and interesting results in a problem that is at the heart of statistics. The real populations that occur in the world are all finite and the infinities that we habitually invoke are constructs, albeit useful but essentially artificial. Another reason for being excited about the work is that it opens the way to further Bayesian studies of the practice of sampling procedures, where several features, and not just one as here, are being investigated.

The authors make much use of the term ‘objective’; what does this mean? My dictionary gives at least two, rather different, meanings: “exterior to the mind” and “aim”. The authors would appear to use both meanings in the same sentence when they say, at the end of Section 1, “A formal objective Bayesian solution . . . is the main objective of this paper”. My opinion is that all statistics is subjective, the subject being the scientist analysing the data, so that the contrary position needs clarification. There is also a confusion for me with the term ‘reference prior’, a term that I have queried in earlier discussions.

An unstated assumption that R and n are independent, given N , has crept into (2) where $\Pr(R|N)$ should be $\Pr(R|n, N)$. The assumption may not be trivial, as when the sampling procedure is to continue until the first non-conforming element is found. Another assumption made is that N is fixed, despite the fact that, in the example of the tortoises, it is unknown. I would welcome some clarification of the role of the sampling procedure.

Perhaps the most interesting section in the paper is 3, where the use of Jeffreys’s prior (12), or (13), superficially very close to the uniform (roughly $1/2$ a confirmation and $1/2$ non-confirmation) gives such different results from it. For example (20) can be written $\pi_r(E_n) = \frac{2n+1}{2n+2}$. Thus Jeffreys gives the same result as Laplace but for *twice* the sample size. Again in the hierarchical model $\pi_r(\text{All} + |n, N)$ is about $\sqrt{n/N}$, equation (13), whereas with the uniform it is about n/N , equation (9), the larger value presumably being due to the prior on θ attaching higher probability than the uniform to values near 1. We therefore have the unexpected situation where an apparently small change in the prior results in an apparently large change in at least some aspects of the posterior.

There are many issues here that merit further study and we should be grateful to the authors for the stimulus to employ their original ideas to do this.

Dennis V. Lindley

Royal Statistical Society’s Guy Medal in Gold in 2002.

University College London, UK

ThomBayes@aol.com

Recibido / Received: 24 de febrero de 2009.

These comments refer to the paper of James O. Berger, José M. Bernardo and Dongchu Sun, (2009). **Natural Induction: An Objective Bayesian Approach**, *Rev. R. Acad. Cien. Serie A. Mat.*, 103(1), 125–135.

© 2009 Real Academia de Ciencias, España.

Comments on: Natural Induction: An Objective Bayesian Approach

Brunero Liseo

I really enjoyed reading the paper. It shed new and clear light to some issues which stand at the core of the statistical reasoning.

In standard statistical models, where the parameter space is a subset of \mathbb{R}^k for some integer k , *reference priors*, and to some extent, *Jeffreys' priors*, offer a way to find a compromise between Bayesian and classical goals of statistics. Usually such a solution lies at the boundary of the Bayesian world, i.e. the objective priors to be used in order to get *good* frequentist behaviour are in general improper. A remarkable exception is however the objective prior for the probability of success θ in a sequence of Bernoulli trials.

Finite populations problems can hardly be approximated by an “infinite population” scenario and, apart from the computational burden, difficulties arise in figuring out what the “boundary of the Bayesian world” would be in these situations. In other words, it is not clear whether a compromise between Bayesian and frequentist procedures is at all possible in finite populations. This paper is then welcome in providing some evidence that, at least, an objective Bayesian analysis of such class of problems is indeed meaningful.

In the rest of the discussion I will focus on the *Law of natural induction*, that is how to evaluate the probability that all the N elements in a population are conforming, given that all the n elements in the sample are. Let R be the unknown number of conforming elements in the population.

The Authors criticize the use of a uniform prior for R and argue that a version of the reference prior, based on the idea of embedding (Berger, Bernardo and Sun, 2009 ([1])), provides more appropriate results. I agree with this conclusion, although the differences are not dramatic. Both uniform and reference priors for R are “symmetric around $N/2$ ”; besides that, the hypothesis $R = N$ does not play a special role: for instance, the two hypotheses $R = N$ and $R = 0$ are given the same weight under both priors; also the cases $R = N$ and $R = N - 1$ have approximately the same prior (and posterior...) probability both under the uniform and the reference prior. These conclusions are perfectly reasonable for an estimation problem when no prior information on R is available. However, the Authors argue that the small value of $\Pr(\text{All} + |n, N)$ “clearly conflicts with the common perception from scientists that, as n increases, $\Pr(\text{All} + |n, N)$ should converge to one, whatever the value of N might be”. This is the crucial point and brings into the discussion the role of models in Statistics. The uniform and the reference prior approaches are not able to catch the idea that $R = N$ and R close to N may be two dramatically different descriptions of the phenomenon: if we are interested in the number of individuals in a population which do not show a genetic mutation, $R = N$ would imply the absence of the mutation with completely different scientific implications from those related to any other value of R .

If the hypothesis $R = N$ has a “physical meaning” then I would have no doubt that the correct analysis to perform is the one described in Section 4. This analysis would make Jeffreys and other objective

Recibido / Received: 24 de marzo de 2009.

These comments refer to the paper of James O. Berger, José M. Bernardo and Dongchu Sun, (2009). **Natural Induction: An Objective Bayesian Approach**, *Rev. R. Acad. Cien. Serie A. Mat.*, 103(1), 125–135.

© 2009 Real Academia de Ciencias, España.

Bayesians happy, since it clearly distinguishes between the statistical “meaning” of the hypothesis $R = N$ and the meaning of other hypotheses, such as $R = 0$ or $R = N - 1$.

In such a situation, formula (28) (or 26) seems a perfectly reasonable “objective Bayesian” answer to the Law of Natural Induction: it is monotonically increasing in n for fixed N and monotonically decreasing in N for fixed n .

So the final question is: can we consider all the scientific questions equivalent to those leading to formula (28)? Should not we take into account the *common perception from scientists* as a guide to choice the best statistical formalization of the problem? To make the point, what happens if a reasonable working model in a specific application, is of the type “ R close to N ”? This is not an infrequent situation; consider, for example, surveys on human or animal populations in order to detect the presence of rare events. In such cases, strong prior information about R might be available and one would rather prefer to perform a reference analysis conditional on some partial prior information, along the lines of Sun and Berger (1998 ([2]), Reference priors with partial information, *Biometrika* [2]).

References

- [1] BERGER, J., BERNARDO, J. M. AND SUN, D. (2009b). Reference priors for discrete parameter spaces. *Submitted*.
- [2] SUN, D. AND BERGER, J. O. (1998). Reference priors with partial information. *Biometrika*, **85**, 55–71.

Brunero Liseo

Dipartimento di studi geoeconomici, linguistici, statistici
e storici per l’analisi regionale
Sapienza Università di Roma,
Italy
brunero.liseo@uniroma1.it

Comments on: Natural Induction: An Objective Bayesian Approach

Kalyan Raman

1 Introduction

Berger, Bernardo and Sun's thought-provoking paper offers a Bayesian resolution to the difficult philosophical problem raised by inductive inference. In a nutshell, the philosophical problem plaguing inductive inference is that no finite number of past occurrences of an event can prove its continuing occurrence in the future. It is thus natural to seek probabilistic reassurance for our instinctive feeling that an event repeatedly observed in the past must be more likely to recur than an event that happened only infrequently. Consequently, as the authors note, the "rule of succession" and the "natural law of induction" have engaged the attention of philosophers, scientists, mathematicians and statisticians for centuries. And rightly so because—despite philosophical qualms about induction—science cannot progress without inductive inferences. The vintage of the induction problem testifies to its difficulty and the pervasiveness of inductive inferences in science reinforces our ongoing efforts to strengthen its underlying logic and fortify its foundations through statistical reasoning. These circumstances necessitate diverse approaches to establish a rigorously justifiable framework for inductive inference.

Berger et al. have made a sophisticated contribution to the literature on rigorously justifying inductive inference, and they have innovatively illuminated an illustrious path blazed by none other than Laplace himself. At the risk of appearing mean-spirited, my main complaint with their solution is the technical virtuosity demanded by their methodology. The mathematical complexities of finding a reference prior are daunting enough to dissuade all but the most lion-hearted in venturing on the search. Given the importance of the problem that Berger et al. address, it may be worthwhile to dredge up an existing solution that seems to be unknown in the statistics literature. In that spirit, I will discuss an alternative approach that produces one of the key results that Berger et al. derive through their reference prior. My approach has the merit of being considerably simpler and more flexible at the expense of possibly not satisfying all the four desiderata listed in Bernardo (2005) ([2]) for objective posteriors, but it does quickly produce a central result in Berger et al. and offers insights into the value of additional replications—an issue that lies at the heart of inductive inference and scientific inquiry. First a few thoughts on the relevance of replications to the topic at hand.

Recibido / Received: 10 de marzo de 2009.

These comments refer to the paper of James O. Berger, José M. Bernardo and Dongchu Sun, (2009). **Natural Induction: An Objective Bayesian Approach**, *Rev. R. Acad. Cien. Serie A. Mat.*, 103(1), 125–135.

© 2009 Real Academia de Ciencias, España.

2 Inductive inference and replications

Bernardo (1979) ([3]) defines a reference posterior in terms of limiting operations carried out on the amount of information about the unknown parameter, obtained from successive independent *replications* of an experiment. Bernardo's definition of reference priors through replications resonates well with a key guiding principle of good scientific research. Replications are the heart and soul of rigorous scientific work—findings that are replicated independently by investigators increase our confidence in the results (Cohen 1990 ([4])). Thus, replications play a fundamental role both in the mathematical definition of a reference posterior and in the scientific process. Clearly, replications are intimately related to inductive inference. It would thus seem conceptually attractive, if, as a by-product of modifying the Laplace Rule of Succession to strengthen its logical basis, we are also able to figure out the optimal informational role of replications.

3 Improving the Laplace rule of succession

Using a reference prior: The solution proposed by Berger et al. to the limitations of the Laplace Rule of succession is displayed in equations (20) and (27) of their paper. Using their notation, the authors' result is that:

$$\pi_u(E_n) = \frac{n + 1/2}{n + 1} \quad (1)$$

which yields faster convergence to unity than the Laplace Rule. The Laplace Rule yields the probability $\pi_u(E_n) = \frac{n+1}{n+2}$. To obtain equation (1), Berger et al. use a hypergeometric model (equation (4) in their paper) together with the reference prior shown in equation (13) of their paper. Equation (13) is obtained by using the Jeffreys prior (equation (12) in Berger et al.) in conjunction with an asymptotic argument which is justified on the basis of exchangeability, as the authors have shown elsewhere. Their logic is sophisticated and beautiful but the price paid for such beauty is that the resultant derivations are arduous. Indeed, Berger and Bernardo (1992) ([1]) themselves admit that the general reference prior method "is typically very hard to implement." Under these circumstances, perhaps the search for a simpler approach is defensible and meritorious of some attention.

Using a beta prior: In Raman (1994) ([7]), I show that the following rule of succession generalizes the Laplace Rule. Suppose that p is the probability that a scientific theory is true, and assume that the prior for p is $\text{Be}(p | \alpha, \beta)$; if we subsequently obtain ' n ' confirmations of the theory, then, using the notation $b_n(E_n)$ to suggest its beta-binomial roots, the probability of observing an additional confirmation is given by,

$$b_n(E_n) = \frac{\alpha + n}{\alpha + \beta + n} \quad (2)$$

Equation (2) follows easily from a result in DeGroot (1975 ([5]), p. 265) guaranteeing equivalence of the sequential updating of $\text{Be}(p | \alpha, \beta)$ with the updating of $\text{Be}(p | \alpha, \beta)$, conditional on having observed " n " successes. The Jeffreys prior $f(p) = \frac{1}{\pi} \frac{1}{\sqrt{p(1-p)}}$, $0 < p < 1$, is a special case resulting from the choice $\alpha = \beta = \frac{1}{2}$ in the prior $\text{Be}(p | \alpha, \beta)$. For that choice of prior, equation (2) reduces to the equation (20) of the Berger et al. paper:

$$\text{For } \alpha = \beta = 1/2, \quad b_n(E_n) = \frac{n + 1/2}{n + 1} \quad (3)$$

Polya (1954) ([6]) recommends a number of properties that an "induction-justifying" rule ought to have—and the beta-binomial rule (equation (2) above) exhibits those desiderata.

Using a general prior, not necessarily beta: It would be natural to object that the above derivation is driven by a specific prior—the Beta distribution. However, in Raman (2000) ([8]), I show that a

generalized rule of succession can be obtained for a general class of priors which includes the Beta distribution as a special case. The generalized rule of succession includes as special cases, the original Laplace Rule, the Beta-Binomial rule and the rule derived in Berger et al. through a reference prior. The exact result is the following: if $g(p)$ is a prior density function with a convergent Maclaurin series representation $g(p) \sim \sum_{i \geq 0} a_i p^i$, then, using the notation g_n to denote the rule of succession under this general prior density,

$$g_n = \sum_{i \geq 0} a_i \frac{i + 1 + n}{i + 2 + n} \quad (4)$$

As special cases, $a_0 = 1, a_i = 0, i \geq 1$, yields the Laplace rule of succession, the choice of a_i as the coefficients in a power-series expansion of $\text{Be}(p | \alpha, \beta)$ results in the beta-binomial rule, which includes, as a special case, the rule of succession for the Jeffreys' prior derived in Berger et al. through a reference prior. Clearly, g_n may be viewed as a linear combination of beta-binomial rules of succession or, with equal right, as a linear combination of Laplacian rules of succession.

From an applied perspective, the Beta density's flexibility and tractability make it an attractive choice for a prior; from a theoretical perspective, the above results show that it suffices for the purpose of generating a more plausible rule of succession than the Laplacian rule, and, in fact, yields results that are identical to Berger et al. Finally, although I do not delve into the topic here, the Beta prior permits derivation of an adaptive controller that shows the value of performing an additional replication as a function of our prior beliefs about the theory, the accumulated evidence in favor of the theory, the precision deemed necessary and the cost of the replication (Raman 1994) ([7]).

Using the Jeffreys' reference prior in Berger et al.: I should remark on the following property of the Jeffreys' reference prior which appears somewhat odd to me. When $N = 1$, it assigns a probability of 0.50, for R , which makes sense. Furthermore, as $N \rightarrow \infty$, the probability $\pi_r(R | N)$ for $R = N$, tends to 0—a result which is attractive. However as N increases, at intermediate values of N , the behavior of $\pi_r(R | N)$ is somewhat odd for $R = N$. Let me explain.

Consider equation (13) in Berger et al.

$$\pi_r(R | N) = \frac{1}{\pi} \frac{\Gamma(R + \frac{1}{2}) \Gamma(N - R + \frac{1}{2})}{\Gamma(R + 1) \Gamma(N - R + 1)}, \quad R \in \{0, 1, \dots, N\}, \quad (13)$$

so $R = N$ implies

$$\pi_r(R | N) = \frac{1}{\pi} \frac{\Gamma(N + \frac{1}{2}) \Gamma(\frac{1}{2})}{\Gamma(N + 1)}.$$

Consider the behavior of the above function as N grows large. The first derivative of $\pi_r(N | N)$ is a complicated expression involving the polygamma function, but if we plot $\pi_r(N | N)$ as a function of ' N ', then we obtain insights. Plotting the function in Mathematica as a function of N (see Figure 1), we find that $\pi_r(N | N)$ at first drops very steeply but that the rate of decline slows down dramatically for $N > 20$. For example, for $100 \leq N \leq 200$, the probability drops from 0.056 at $N = 100$ to 0.039 at $N = 200$.

Thus $\pi_r(N | N)$ is insensitive to new information for large but finite values of N , which is the case that would be of greatest pragmatic interest in scientific theory-testing. It would be useful if the authors could comment on the significance of this property for natural induction.

4 Conclusion

My thoughts on the elegant analysis of Berger et al. are driven by an entirely applied perspective. Consequently, I seek the most parsimonious and mathematically tractable route to model-building. The alternative approach I have described lacks the technical sophistication and mathematical rigor of the authors' reference prior approach—its primary justification is its ease of use and pliability at addressing a broader set

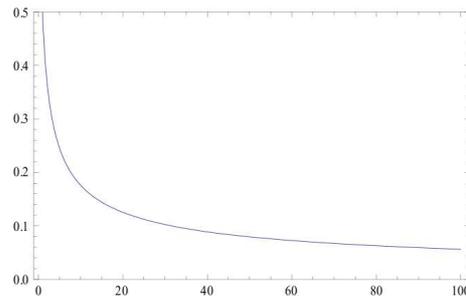


Figure 1. $\pi_r(N | N)$ as a function of N .

of issues (such as the development of an optimal controller to balance the tradeoffs involved in replicating experiments). I realize that these broader issues are not necessarily relevant to the authors—but even so, I would argue that the authors may benefit from thinking about how reference priors can address these questions better than my naïve approach based on a mathematically convenient family of conjugate priors, because their reflection on the applied concerns I have raised could lead to new results that would broaden the scope and scientific impact of reference priors on researchers across multiple disciplines.

In conclusion, I applaud the authors for their innovative application of a powerful new technique to an important and vexing problem of ancient vintage, and hope that some of their future work on reference priors makes the methodology less mysterious, thereby disseminating their ideas to a wider audience and paving the way for new applications based on reference priors.

References

- [1] BERGER, J. O. AND BERNARDO, J. M., (1992). On the development of reference priors. in *Bayesian Statistics*, **4**. (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 35–60 (with discussion).
- [2] BERNARDO, J. M., (2005). Reference analysis, in *Handbook of Statistics*, **25**, 17–90. D. K. Dey and C. R. Rao, (eds.) Amsterdam: Elsevier.
- [3] BERNARDO, J. M., (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B*, **41**, 113–147 (with discussion). Reprinted in *Bayesian Inference*, **1** (G. C. Tiao and N. G. Polson, eds.) Oxford: Edward Elgar, 229–263.
- [4] COHEN, JACOB, (1990). Things I have Learned So Far, *American Psychologist*, **45**, (December), 1304–1312.
- [5] DEGROOT, MORRIS H., (1975). *Probability and Statistics*, Reading, MA: Addison-Wesley.
- [6] POLYA, GEORGE, (1968). *Mathematics and Plausible Reasoning*, Vol. **2**, Princeton, NJ: Princeton University Press.
- [7] RAMAN, KALYAN, (1994). Inductive Inference And Replications: A Bayesian Perspective, *Journal of Consumer Research*, March, **20**, 633–643.
- [8] RAMAN, KALYAN, (2000). The Laplace Rule of Succession Under A General Prior, *Interstat*, June, **1**, <http://interstat.stat.vt.edu/interstat/articles/2000/abstracts/u00001.html-ssi>.

Kalyan Raman
 Medill IMC Department,
 Northwestern University,
 USA
 k-raman@northwestern.edu

Comments on: Natural Induction: An Objective Bayesian Approach

Christian P. Robert

This is a quite welcomed addition to the multifaceted literature on this topic of natural induction that keeps attracting philosophers and epistemologists as much as statisticians. The authors are to be congratulated on their ability to reformulate the problem in a new light that makes the law of natural induction more compatible with the law of succession. Their approach furthermore emphasize the model choice nature of the problem.

First, I have always been intrigued by the amount of attention paid to a problem which, while being formally close to Bayes' own original problem of the binomial posterior, did seem quite restricted in scope. Indeed, the fact that the population size N is supposed to be known is a strong deterrent to see the problem as realistic, as shown by the (neat!) Galapagos example. My first question is then to wonder how the derivation of the reference prior by Berger, Bernardo, and Sun extends to the case when N is random, in a rudimentary capture-recapture setting. An intuitive choice for $\pi_r(N)$ is $1/N$ (since N appears as a scale parameter), but is

$$\frac{\binom{R}{r} \binom{N-R}{n-r}}{\binom{N}{n}} \frac{\Gamma(R+1/2)\Gamma(N-R+1/2)}{R!(N-R)!} \frac{1}{N}$$

summable in both R and N ? (Obviously, impropriety of the posterior does not occur for a fixed N .)

As exposed in the paper, one reason for this special focus on natural induction may be that it leads to such a different outcome when compared with the binomial situation. Another reason is certainly that Laplace succession's rule seems to summarise in the simplest possible problem the most intriguing nature of inference. And to attract its detractors, from the classical Hume's (1748) ([1]) to the trendy Taleb's (2007) ([2]) "black swan" argument (which is not the issue here, since the "black swan" criticism deals with the possibility of model changes).

Second, the solution adopted in the paper follows Jeffreys' approach and I find this perspective quite meaningful for the problem at hand. Indeed, while N can be seen as $(N-1)+1$, i.e. as one of the $N+1$ possible values for R , the consequence of having R equal to either N or 0 lead to atomic distributions for the number of successes. Thus, to distinguish those two values from the other makes sense even outside a testing perspective. In Jeffreys' (1939) original formulation, both extreme values, 0 and N , are kept separate, with a prior probability k between $1/3$ and $1/2$. I thus wonder why the authors moved away from this original perspective. The computation for this scenario does not seem much harder since $\pi_r(0|N) = f(N)$ as well and the equivalent of (22) would then be

Recibido / Received: 11 de marzo de 2009.

These comments refer to the paper of James O. Berger, José M. Bernardo and Dongchu Sun, (2009). **Natural Induction: An Objective Bayesian Approach**, *Rev. R. Acad. Cien. Serie A. Mat.*, 103(1), 125–135.

© 2009 Real Academia de Ciencias, España.

$$\pi_{\phi}(\text{All} + |n, N) = \left(1 + \frac{k}{1-2k} \frac{f(n) - 2f(N)}{1-2f(N)}\right)^{-1},$$

which is then $(1 + 0.5f(n))^{-1}$ for N large. In this case, (24) is replaced with $\sqrt{n}/(\sqrt{n} + 2/\sqrt{\pi})$, not a considerable difference.

In conclusion, I enjoyed very much reading this convincing analysis of a hard “simple problem”! It is unlikely to close the lid on the debate surrounding the problem, especially by those more interested in the philosophic side of it, but rephrasing natural induction as a model choice issue and advertising the relevance of Jeffreys’ approach to this very problem have bearings beyond the “simple” hypergeometric model.

References

- [1] HUME, D., (1748). *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*, (2000 version: Oxford University Press).
- [2] TALEB, N. N., (2007). *The Black Swan: The Impact of the Highly Improbable*. Penguin Books, London.

Christian P. Robert
CEREMADE
Université Paris Dauphine,
France
xian@ceremade.dauphine.fr

Comments on: Natural Induction: An Objective Bayesian Approach

Raúl Rueda

According to the authors: “The conventional use of a uniform prior for discrete parameters can ignore the structure of the problem under consideration”. This motivates the introduction of a hierarchical structure

$$p(R|N) = \int_0^1 \text{Bi}(R|N, \theta) \text{Be}(\theta|1/2, 1/2) d\theta,$$

where $\text{Bi}(R|N, \theta)$ is the binomial distribution with parameters (N, θ) and $\text{Be}(\theta|1/2, 1/2)$ is the reference prior for θ in this case.

However, it must be pointed out that the assumption of exchangeability to justify the hierarchical structure is also valid for the uniform prior, by replacing the $\text{Be}(\theta|1/2, 1/2)$ distribution with a uniform in $(0, 1)$ yielding as a prior $p[R] = 1/(N + 1)$.

Anyway, the reference *Rule of Succession* is essentially the same as Laplace’s, but there is a difference in $\pi[\text{All} + |n, N]$ when n is small compared with N . This difference disappears when $n \rightarrow N$.

Even though the authors find equations (9) and (11) to be “dramatically different”, suggesting a contradictory behaviour, this is perfectly possible for example, in the case of rare events, such a finding a person who suffers from a disease with a prevalence of one in a million. If the conforming event is the absence of the disease, then there is a high probability that we observe another conforming element given that all elements in the sample are conforming. At the same time, the probability that all are conforming is close to zero, so in this case (9) and (11) are both valid, and the behaviour of (22) becomes more difficult to accept.

Raúl Rueda

Departamento de Probabilidad y Estadística,
Universidad Autónoma Nacional de México,
México

pinky@sigma.iimas.unam.mx

Recibido / Received: 10 de marzo de 2009.

These comments refer to the paper of James O. Berger, José M. Bernardo and Dongchu Sun, (2009). **Natural Induction: An Objective Bayesian Approach**, *Rev. R. Acad. Cien. Serie A. Mat.*, 103(1), 125–135.

© 2009 Real Academia de Ciencias, España.

Comments on: Natural Induction: An Objective Bayesian Approach

S. L. Zabell

Berger, Bernardo, and Sun (BBS) briefly allude to exchangeability in their paper. Since I personally find this the most natural way to view these types of questions, I begin by discussing their results from this point of view.

Suppose X_1, \dots, X_N is a finite exchangeable sequence of 0s and 1s with respect to a probability P . The simplest such probability assignment is the one corresponding to an urn $U_{R,N}$ with N balls, R 1s and $N - R$ 0s, where we draw the balls out at random without replacement. Denote this hypergeometric probability assignment by $H_{R,N}$. If $S_N = X_1 + \dots + X_N$ and $p_R = P(S_N = R)$, then by exchangeability $P = \sum_R p_R H_{R,N}$. Call this the finite de Finetti representation.

In general a finite exchangeable sequence X_1, \dots, X_N cannot be extended (and remain exchangeable), but if it can be indefinitely extended then it admits an integral representation: for some probability measure Q on the unit interval one has

$$P(S_N = R) = \int_0^1 \binom{N}{R} p^R (1-p)^{N-R} dQ(p);$$

this is the celebrated de Finetti representation theorem.

Thus P can be thought of arising in two apparently different, but actually stochastically equivalent ways: 1) choose a p -coin according to Q , and toss it N times, or 2) choose an urn $U_{R,N}$ according to p_R , and then draw the balls out at random from it without replacement.

In Laplace's famous 1774 paper he took the first route, adopting the flat $dQ(p) = dp$. This special prior has, as BBS note, the interesting properties that $P(S_N = R) = 1/(N + 1)$, for $0 \leq R \leq N$; and for any $n < N$, $P(X_{n+1} = 1 | S_n = r) = (r + 1)/(n + 2)$; the "rule of succession".

The classical Laplacean analysis raises a number of questions: the nature of p (presumably some form of "physical probability"); the implicit presence of an (at least in principle) infinitely extendable sequence; and exactly what is meant by Laplace's idea of sampling with replacement from an infinite population. So it was perhaps inevitable that someone would eventually ask about the corresponding state of affairs if you sample without replacement from a finite population and make the natural assumption that all possible fractions of 0s and 1s are equally likely.

This is what C. D. Broad did in 1918. But as the Bible tells us, "there is nothing new under the sun". The analysis of sampling without replacement from a finite population, using a uniform prior on the fraction of "conformable elements", had already been carried out more than a century earlier, by Prevost and L'Huilier in 1797! In their direct attack on the problem it is necessary to prove a not entirely trivial combinatorial identity in order to establish the rule of succession; see Todhunter (1865 ([2]), pp. 454–457), Zabell (1988).

Recibido / Received: 6 de marzo de 2009.

These comments refer to the paper of James O. Berger, José M. Bernardo and Dongchu Sun, (2009). **Natural Induction: An Objective Bayesian Approach**, *Rev. R. Acad. Cien. Serie A. Mat.*, 103(1), 125–135.

© 2009 Real Academia de Ciencias, España.

The result was a surprise: as Prevost and L'Huilier note, the rule of succession $(r+1)/(n+2)$ (for a sample of size n from a population of size N) does not depend on N , and is the same of Laplace's! As Todhunter remarks, "The coincidence of the results obtained on the two different hypotheses is remarkable" (1865, p. 457).

But in fact it is not remarkable; from the perspective of the finite de Finetti representation one does not need to evaluate a tricky combinatorial sum, nor is the coincidence of the rules of succession in any way surprising. A finite exchangeable sequence X_1, \dots, X_n is completely characterized by the probabilities $p_R = P(S_N = R)$; and so whether the process is generated by first picking p at random from the unit interval and then tossing a p -coin N times, or by randomly selecting an urn $U_{R,N}$ and then drawing its balls out one at a time without replacement, *you get stochastically identical processes* (because in both cases $P(S_N = R) = 1/(1+N)$); it's not just that the rules of succession coincide but *everything* is the same!

It is interesting to trace intellectual dependencies. Broad (1924 ([1]), Section 3) attributes his "interest in the problems of probability and induction" to W. E. Johnson (a Cambridge logician who derived Carnap's "continuum of inductive methods" more than 20 years before Carnap did); and Broad's 1918 analysis was in turn an important influence on Sir Harold Jeffreys, who tells us that "It was the profound analysis in this paper that led to the work of Wrinch and myself" (Jeffreys, 1961, p. 128).

One of the reasons Broad's paper made such a splash at the time was his noting that although (under the uniform prior) the probability that the next crow will be black, given all n crows to date have been black, is nearly one if n is large, $(n+1)/(n+2)$, the probability that *all* crows (in the finite population of N) are black, given the n so far are black, is small for $n \ll N$, namely $(n+1)/(N+2)$.

This property was seen as a problem for any attempt at a mathematical explication of induction, and led Wrinch and Jeffreys to write their papers. As BBS note, in Section 3.2 of his book Jeffreys makes the natural suggestion to allocate some initial probability *independent of N* to natural laws. BBS say "The simplest choice is to let $Pr(R = N) = 1/2$ "; but as far as I can tell, Jeffreys usually puts the cases $R = N$ and $R = 0$ on an equal footing. So I would have liked to have seen some further discussion of this suggestion, which clearly treats the cases asymmetrically (since of course if $P(R = N) = P(R = 0) = 1/2$, this would account for all the probability).

There is, however, an interesting historical precedent for viewing matters from such an asymmetric perspective. The Reverend Dr. Richard Price, in his discussion at the end of Bayes's famous essay (Bayes, 1764), considers the application of Bayes's results to the problem of induction. Bayes's version of the rule of succession is different from Laplace's (Bayes's rule is $1 - 2^{-(n+1)}$, a different answer to a different question), but the point here is when Price thinks one should start counting. Bayes's results, he tells us, apply to

[A]n event about the probability of which, antecedently to trials, we know nothing, that it has happened *once*, and that it is enquired what conclusion we may draw from hence with respect to the probability of it's happening on a second trial.

Note the requirement that the event will have already occurred *once*. Why? Imagining "a solid or die or whose number of sides and constitution we know nothing", Price explains:

The first throw only shews that *it has* the side then thrown It will appear, therefore, that *after* the first throw and not before, we should be in the circumstances required by the conditions of the present problem, and that the whole effect of this throw would be to bring us into these circumstances. That is: the turning the side first thrown in any subsequent single trial would be an event about the probability or improbability of which we could form no judgment, and of which we should know no more than that it lay somewhere between nothing and certainty. With the second trial then our calculations must begin

This leads Price to consider the famous (or infamous example?) of the rising of the sun:

Let us imagine to ourselves the case of a person just brought forth into this world and left to collect from his observations the order and course of events what powers and causes take

place in it. The Sun would, probably, be the first object that would engage his attention; but after losing it the first night he would be entirely ignorant whether he should ever see it again. He would therefore be in the condition of a person making a first experiment about an event entirely unknown to him. But let him see a second appearance or one *return* of the Sun, and an expectation would be raised in him of a second return

I take it that one would adopt a reference prior only absent substantial background information. The interest of Price's remarks is they address in a serious way just how, that being the case, epistemic asymmetry might still be natural. So, as indicated earlier, I would have been interested to see further discussion in BBS of when the assignment $P(R = N) = 1/2$ is appropriate. If, for example, I am a doctor trying out a new procedure or drug, then would I not want some "Jeffreys-like" prior probability assigned to both extremes? Is the reference prior assignment of $1/2$ most appropriate in "Price-like" situations?

One final question for BBS. Suppose there are $t \geq 3$ possibilities (say a, b, c) rather than just the two of conforming and non-conforming. Just as seeing all a thus far should increase the probability that all elements of the population are a , so too seeing, say, only a and b but no c should increase the probability that there are no c in the population. In general there are 2^{t-1} sub-simplexes to which one would like to assign some positive probability. What would be the reference prior approach in this case?

In any case I would like to complement the authors on a most interesting and stimulating paper.

References

- [1] BROAD, C. D., (1924). *Philosophical Autobiography*, Contemporary British Philosophy: Personal Statements, First Series (ed. J. H. Muirhead), G. Allen and Unwin, London, 77–100.
- [2] TODHUNTER, I., (1865). *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace*, Macmillan, London.

S. L. Zabell

Departments of Mathematics and Statistics,
Northwestern University,
IL, USA
zabell@math.northwestern.edu

Rejoinder on: Natural Induction: An Objective Bayesian Approach

J. O. Berger, J. M. Bernardo and D. Sun

We appreciate the positive general comments of most of the discussants. And, of course, we are grateful for the interesting and thought-provoking additional insights and comments that they have provided. We provide below a response to these comments.

Girón and Moreno. We certainly agree with Professors Girón and Moreno on the interest in sensitivity of *any* Bayesian result to changes in the prior. That said, we also consider of considerable pragmatic importance to be able to single out a *unique*, particular prior which may reasonably be proposed as the reference prior for the problem under study, in the sense that the corresponding posterior of the quantity of interest could be routinely used in practice when no useful prior information is available or acceptable. This is precisely what we have tried to do for the twin problems of the rule of succession and the law of natural induction.

The discussants consider the limiting binomial version of the Law of Natural Induction, and focused on the version that can be stated in the language of hypothesis testing involving $H_0 \equiv \{p = 1\}$. They then noted that a popular objective Bayesian approach to hypothesis testing is to use intrinsic priors on the alternative, which tend to be more concentrated about the null value than the $\text{Be}(p | 1/2, 1/2)$ prior we use. The notion is that, if a problem is posed as that of testing a ‘privileged’ null hypothesis, then realistic alternatives will tend to be close to the null value, and the prior distribution – even in supposedly objective procedures – should reflect this. Thus a strong case can be made for use of intrinsic priors in that setting.

The natural induction problem, however, is not a problem with a privileged null hypothesis in this sense; there is no a-priori notion that values of p near 1 are more believable than other values. Hence we would argue that, for the natural induction problem, the analysis we propose is the preferred objective procedure.

Lindley. As one would expect from a well known subjectivist, Professor Lindley questions our use of the word *objective*. Although we certainly agree that *any* statistical analysis is really subjective (for the crucial model assumptions are typically made from subjective choices) we use the term objective in the precise sense that the Bayesian result obtained only depends on the model assumed and the data obtained and, therefore, has precisely the same claim to objectivity that is often made of frequentist statistics. Unfortunately, Bayesian methods are still often disregarded because of the completely wrong impression that they *must* use subjective priors; by use of the word objective we simply want to stress that Bayesianism need not be rejected for this reason.

Professor Lindley is certainly right in pointing out that we have assumed that R and n are independent given N , so that a more appropriate notation would have been $\text{Pr}(R | n, N)$ rather than $\text{Pr}(R | N)$. Indeed,

Recibido / Received: 25 de marzo de 2009.

This rejoinder refers to the paper of James O. Berger, José M. Bernardo and Dongchu Sun, (2009). **Natural Induction: An Objective Bayesian Approach**, *Rev. R. Acad. Cien. Serie A. Mat.*, **103**(1), 125–135.

© 2009 Real Academia de Ciencias, España.

an inverse sampling procedure would have produced different results. We have certainly worked conditional on N so that, if N is not really known, a range of results must be quoted, as in the galápagos example. That said, there are many important applications where N is actually known. For instance, in industrial production, one may be interested in the probability that all items in a production of size N are conforming, given that the elements of a random sample of size n are all conforming, and the value of N is typically known in those cases.

As Professor Lindley interestingly points out, our rule of succession gives the same result as Laplace, but for twice the sample size. Considering the practical implications of this fact (for instance when setting insurance policies in the industrial situations described above), the situation is a good example of the possible rewards in careful choice of the reference prior.

Liseo. As Professor Liseo points out, model choice priors (or precise hypothesis testing priors in the language we use in the paper) must be different from those of estimation. Proposed hypothesis typically have a precise physical meaning, and its possible validity has to be reflected in the prior. One may choose to do this subjectively, or one may try an objective approach using a reference prior where the quantity of interest is whether or not the hypothesis is true, as we have tried to do in the problem of natural induction.

Professor Liseo poses the question of what to do if the application suggests a hypothesis of the type “ R is close to N ”. In line with the argument above we would suggest eliciting a value $0 < \alpha < 1$ such that the closest integer to αN is close enough to N for the application in mind, and then use a marginal reference prior with $\Pr(\alpha N \leq R \leq N) = 1/2$, and a conditional reference prior for R given $0 \leq R < \alpha N$ which would be the conditional reference prior (14) renormalized to the set $0 \leq R < \alpha N$.

Raman. Professor Raman is certainly right when he points out the fact that the derivation of reference priors is often less than trivial (although, their *use* is often trivial, once someone has done the relevant research and produced the appropriate answers.) That said, it is certainly interesting to be able to derive some specific instances of the reference analysis results using simpler techniques. For instance, invariance arguments provide simple and powerful techniques to provide objective priors (which are typically found to be the same as those derived from reference analysis), but many important problems do not have the required structure to use them; reference analysis provides a general argument, which does not depend of particular properties of the problem under investigation.

As Professor Raman points out, the continuous version of our rule of succession is the particular case which obtains if a conjugate Beta prior is used, and the particular case $\alpha = \beta = 1/2$ is chosen; he further outlines how this can interestingly be extended to the use of beta mixtures. The argument by Polya provides some additional reasons to choose a conjugate prior in this problem, beyond mathematical expediency, but one would need some special argument to select precisely $\alpha = \beta = 1/2$, if one wanted to argue that $(n + 1/2)/(n + 1)$ is indeed the appropriate rule of succession.

Professor Raman seems surprised by the fast convergence to zero of the reference probability $\pi_r(N | N)$. This is a reflection of the fact that, as N increases, its precise value does not matter much. Besides, reference priors should not really be analyzed as prior probabilities (indeed, in continuous parameter problems those are typically improper) but as technical devices to produce posterior probabilities. And it is this the reference posterior probabilities what should be carefully discussed for any possible data set.

Robert. Professor Robert focuses attention on extending the problem to consider the case where N is unknown, a situation also mentioned by Professor Lindley. The objective Bayesian answer is indeed to specify an objective prior for N and, as Professor Robert points out, the intuitive choice for an objective prior for the scale-like parameter N would be $\pi(N) = 1/N$. Given the time restrictions imposed by the Journal, we have not had time to explore the consequences of this choice, let alone to derive the appropriate conditional reference prior for N , $\pi(N | R)$, when R is the quantity of interest, but we certainly take note of this problem for future research.

As Professor Robert points out, it is not difficult to implement Jeffreys suggestion for the prior probabilities of $R = 0$ and $R = N$, together with the conditional reference prior for $\pi(R | n, N)$ but, if one is to follow the reference analysis approach, one should use a marginal reference prior of the quantity of

interest, which here is whether or not $R = N$, and this leads to $\pi(R = N) = \pi(R \neq N) = 1/2$ rather than to the values which Jeffreys suggested, namely $\pi(R = 0) = \pi(R = N) = k$, with $1/3 \leq k \leq 1/2$. As already raised in the response to Professors Girón and Moreno, the issue is partly one of when and how one should formally admit subjective knowledge into objective procedures. By incorporating into the reference analysis the knowledge that $R = N$ is a clearly important possibility, we went part way to Jeffreys solution. It seemed to us, however, that the contexts in which the Laws are discussed are not contexts in which $R = 0$ has any special believability.

Rueda. Professor Rueda is certainly right when he points out that the exchangeability argument only implies a hierarchical model with a a prior distribution $\pi(\theta)$, which could be the $\text{Be}(\theta | 1/2, 1/2)$ we use, the uniform prior leading to Laplace rule, or any other proper prior for θ . Our particular choice for $\pi(\theta)$ is motivated by the fact that this is *the* reference prior which corresponds to the *implied* integrated model which, as argued in the paper, is the Binomial $\text{Bi}(r | n, \theta)$.

The very different behavior of (9) and (11) —and, correspondingly, (18) and (20)— as the population size N increases is mathematically driven by nature of the priors used, with a prior probability for $R = N$ which depends on N and goes to 0 as N increases. If R is the only parameter of interest, these are indeed the appropriate results for the problems considered. If, however, the quantity of interest is whether or not $R = N$, then a prior probability for $R = N$ which does not depend on N is required, and the results, (25) and (22), are very noticeably different.

Zabell. We are very grateful to Professor Zabell for his authoritative insights into the history of the problems considered in this paper: we were not aware of the many precedents to Broad's derivation of the Laplace rule of succession for the finite case. We are similarly not aware of any previous derivation of the reference rule of succession for the finite case given by Equation (19) (a result given by Perks (1947) for the continuous case), but then, we would not be surprised if there is one!

As also noted by Professor Robert, Professor Zabell wonders about the advantages of using a symmetric prior for $\text{Pr}(R = 0)$ and $\text{Pr}(R = N)$. As mentioned in our response to Professor Robert, such decisions are indeed context dependent; there is subjectivity in terms of how one chooses to formalize a reference analysis and, in a situation such as drug testing, we would certainly propose a different formulation than that which we felt to be reasonable for the Laws.

As Professor Zabell correctly points out, one should only use a reference prior in the absence of public, commonly accepted substantial background information. If there is such information, this should be explicitly stated and be made part of the model. Often, this will take the form of a hierarchical model, and then one would need an objective prior for the hyperparameter, possibly the reference prior associated to the corresponding integrated model.

The extension of the results presented to more than two possibilities in the population is a very interesting possible generalization. We expect that the same basic reference analysis techniques could be used for this generalization.

Acknowledgement. We would like to conclude by thanking again our discussants and, last but not least, by thanking Professor Pellicer and the staff of RACSAM for inviting us to present this work, and for processing the paper in a record time.

James O. Berger

Duke University,
USA

berger@stat.duke.edu

José M. Bernardo

Universitat de València,
Spain

jose.m.bernardo@uv.es

Dongchu Sun

University of Missouri-Columbia,
USA

sund@missouri.edu