Bayesian Analysis (2009) 4, Number 1, pp. 151–158

Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters

Frank Tuyl*, Richard Gerlach† and Kerrie Mengersen‡

Abstract. It is argued that the posterior predictive distribution for the binomial and multinomial distributions, when viewed via a hypergeometric-like representation, suggests the uniform prior on the parameters for these models. The argument is supported by studying variations on an example by Fisher, and complements Bayes' original argument for a uniform prior predictive distribution for the binomial. The fact that both arguments lead to invariance under transformation is also discussed.

Keywords: Bayesian inference; binomial distribution; invariance; noninformative priors; Jeffreys prior

A Introduction

Interval estimation of the binomial parameter θ, representing the true probability of a success, is a problem of long standing in statistical inference. In the case of prior ignorance, or no knowledge about θ, a noninformative prior can be used to 'let the data speak for themselves'. Such a prior has been called a consensus prior and aims to be suitable as a standard for scientific communication (Bernardo 2005). However, there is some debate as to the optimal choice for the binomial parameter. Bernardo and Ramon (1998) stated a nice desideratum for a noninformative prior: "Proper or improper, what must be required from non-subjective priors is that, for any data set, they lead to sensible, data-dominated, *posterior* distributions." Four 'plausible' noninformative priors were listed by Berger (1985, p.89): the Bayes-Laplace beta$(1, 1)$, the Jeffreys/reference beta$(\frac{1}{2}, \frac{1}{2})$, the Haldane beta$(0, 0)$ and Zellner's prior (which is U-shaped but not a beta density).

Bayes (1763) applied the uniform prior to derive the beta posterior that is the normalised binomial likelihood function. It is not well known that Bayes favored this prior as a result of considering the observable random variable x as opposed to the unknown parameter θ, which is an important difference (Edwards 1978; Stigler 1982). In modern terms, what Bayes clearly proposed as a reasonable representation of prior ignorance (p.392: "...that concerning such an event I have no reason to think that, in

*Hunter New England Population Health, Newcastle, Australia, mailto:frank.tuyl@newcastle.edu.au

†Faculty of Economics and Business, University of Sydney, Australia, mailto:R.Gerlach@econ.usyd.edu.au

‡School of Mathematical Sciences, Queensland University of Technology, Australia, mailto:k.mengersen@qut.edu.au

 DOI:10.1214/09-BA405

a certain number of trials, it should rather happen any one possible number of times than another."), is a uniform prior predictive distribution

$$p(x) = \int_0^1 p(\theta)p(x|\theta)d\theta = \frac{1}{n+1}, \quad x = 0, \ldots, n,$$

from which $p(\theta) = 1$ follows under mild conditions. For further details, see Stigler (1982) who emphasised that consequently Bayes did *not* appeal to the 'suspect principle of insufficient reason', i.e. the notion that lack of knowledge about a parameter implies a uniform prior.

Given actual data x, the *posterior* predictive distribution of an unobserved quantity y is

$$p(y|x) = \int p(y|\theta)p(\theta|x)d\theta,$$

which is a powerful Bayesian tool for forecasting purposes. Rather than, for example, 'plugging in' the maximum likelihood estimate, integration is applied across all possible values of θ, thereby incorporating parameter uncertainty.

The aim of this paper is to show that the posterior predictive distributions for the binomial and multinomial models also suggest the uniform beta$(1, 1)$ or Bayes-Laplace (B-L) prior as the natural representation of prior ignorance, and thus a suitable consensus prior. Geisser (1984) already gave a convincing argument in favor of the uniform prior based on hypergeometric sampling. This clearly supports, but is different from, the argument presented here. There is no denying, however, that the current objective Bayesian choice is the Jeffreys/reference beta$(\frac{1}{2}, \frac{1}{2})$ prior (Box and Tiao 1973; Bernardo and Smith 1994), so our comparisons focus on these two priors.

The main arguments are given in Section 2, supported by examples derived from Fisher (1973). Invariance of the B-L prior is discussed in Section 3.

B Main arguments

A hypergeometric-like representation of the posterior predictive distribution is given for the binomial model, followed by the generalisation to the multinomial case.

B.1 Binomial model

For the binomial model the posterior predictive distribution describes the probability of y successes in m trials given x successes in n trials. The general beta(a, b) prior leads to the beta-binomial distribution:

$$p(y|m, x, n) = \binom{m}{y} \frac{\Gamma(n+a+b)\Gamma(y+x+a)\Gamma(m+n-y-x+b)}{\Gamma(x+a)\Gamma(n-x+b)\Gamma(m+n+a+b)}.$$

Using factorials instead of gamma functions, Thatcher (1964) referred to y as having a negative hypergeometric distribution, presumably based on the re-arrangement (not

given by Thatcher)

$$p(y|m,x,n) = \frac{\binom{y+x+a-1}{y}\binom{m+n-y-x+b-1}{m-y}}{\binom{m+n+a+b-1}{m}}, \tag{1}$$

which generalises the expressions by Geisser (1993, p.52) and Jaynes (2003, p.563), which are for $a = b = 1$ only; Geisser (1984) gave expressions for $a = b = \frac{1}{2}$ and $(a = 0, b = \frac{1}{2})$.

Similar to the negative binomial, the negative hypergeometric representation of the posterior predictive distribution of y given by (1) can be seen as the product of two probabilities:

$$p(y|m,x,n) = \frac{n+a+b-1}{m+n+a+b-1}\frac{\binom{y+x+a-1}{y}\binom{m+n-y-x+b-1}{m-y}}{\binom{m+n+a+b-2}{m}}. \tag{2}$$

This representation, which, as far as we know, has not appeared before, gives an interesting view of the beta-binomial. It also implies that, for integer values of a and b, it can be evaluated easily in any software packages that provide the hypergeometric distribution.

The hypergeometric-like component of (2) suggests the uniform prior $(a = b = 1)$, leading to symmetry in y and x. Applying this prior the resulting multiplier $(n + 1)/(m+n+1)$ is sensible as well: for example, the ratio of $p(y|m,x,n)$ and $p(x|n,y,m)$ is $(n+1)/(m+1)$, which simply reflects the fact that y and x have $m+1$ and $n+1$ possible values, respectively. This ratio is 1 when $m = n$ so that $p(y|n,x,n)$ *equals* $p(x|n,y,n)$; as illustrated below, this is eminently reasonable, but follows from the uniform prior only.

We support this argument with an example by Fisher (1973, p.137) who stated, "The likelihood of observing 14 successes out of 21 as judged by data showing 3 successes out of 19, is exactly the same as the likelihood in prospect of observing 3 successes out of 19, judged on the basis of experience of 14 successes out of 21." (In the same section Fisher emphasised that likelihood is different from probability.) While it is tempting to interpret the two statements as conditional, this kind of equality only eventuates from Fisher's unconditional approach here: based on (2), conditional (i.e. posterior) probabilities such as these are not generally the same when n and m are not the same; this follows immediately from considering, for example, n large and m very small.

As referred to in the Introduction, the B-L beta$(1,1)$ and Jeffreys beta$(\frac{1}{2},\frac{1}{2})$ priors are two of a set of four plausible priors (Berger 1985, p.89). The other two are the Zellner and Haldane priors. The Zellner prior is proportional to $\theta^{\theta}(1-\theta)^{1-\theta}$ and is proper with normalisation constant 1.6186, which is also the pdf value at the extremes, making it a less extreme U-shape than the Jeffreys prior. The Haldane prior is beta$(0,0)$ and

improper, so that for $x = 0$ or $x = n$ the posterior predictive (1) or (2) is not defined. This follows from the impropriety of the posterior distribution, which, as Bernardo (1979) noted, is less than adequate.

In Table B.1 the posterior predictive probabilities induced by these four priors are compared for 11 scenarios, comprising Fisher's scenario (Case 1) and a variation on it (Case 2), followed by some extreme data outcomes (Cases 3 and 4) and a simple example with $m = n$ (Case 5).

Case	y	m	x	n	beta$(1,1)$ Bayes-Laplace	beta$(\frac{1}{2},\frac{1}{2})$ Jeffreys	beta$(0,0)$ Haldane	$\propto \theta^{\theta}(1-\theta)^{(1-\theta)}$ Zellner
$1a$	14	21	3	19	0.000619	0.000461	0.000328	0.000500
$1b$	3	19	14	21	0.000681	0.000645	0.000606	0.000649
$2a$	14	20	3	20	0.000255	0.000186	0.000130	0.000204
$2b$	3	20	14	20	0.000255	0.000235	0.000214	0.000238
$3a$	1	1	10	10	0.9167	0.9545	N/A	0.9278
$3b$	10	10	1	1	0.1667	0.3364	N/A	0.2150
$3c$	9	10	1	1	0.1515	0.1602	N/A	0.1673
$4a$	0	1	10	10	0.0833	0.0455	N/A	0.0722
$4b$	10	10	0	1	0.0152	0.0160	N/A	0.0167
$5a$	4	4	2	4	0.1190	0.1289	0.1429	0.1259
$5b$	2	4	4	4	0.1190	0.0663	N/A	0.0979

Table 1: Posterior predictive scenarios: $p(y|m, x, n)$

Although the Case 1 probabilities are quite small, it is easy to see the relative differences. Of course, no general statement can be made about the 'correctness' of the priors based on this scenario alone, though clearly the B-L results are closest to conforming with Fisher's statement. This is due to the scenario Fisher chose, which we slightly adjusted so that $m = n$ (Case 2). As expected, the B-L probabilities are now the same for $2a$ and $2b$, and we argue that 14/20 given 3/20 should indeed be as unexpected as 3/20 given 14/20. In contrast, given a beta(a, a) prior with $a < 1$ (Jeffreys and Haldane), the probability of the latter exceeds that of the former, and the reverse applies when $a > 1$; i.e. if the future event is more 'aligned' with the prior than the past event, in terms of being closer to or further away from an extreme, it has greater probability than the reverse case. This appears to leave $a = 1$ as the only beta prior that 'lets the data speak for themselves'. In the remainder we discuss differences between B-L and Jeffreys only, with the Zellner values typically between the two and the Haldane values mostly non-existent.

Cases 3, 4 and 5 support the argument that the B-L prior leads to more natural posterior predictive probabilities than the Jeffreys prior; the same conclusions are drawn from larger samples. Note that the B-L differences between a and b for Cases 1, 3 and 4 are purely due to the $(n + 1)/(m + 1)$ factor: in Cases 3 and 4, for example, this is $11/2 = 5.5$. As expected, this causes quite a large difference between the B-L a and b scenarios. This difference is smaller under the Jeffreys prior, but these results are difficult to justify; for example, in Case $3b$ the probability of 0.3364, given the evidence of only one observation, is more than 20 times greater than the Case $4b$ probability

based on one failure instead of one success. Moving from Case 3*b* to 3*c* there is a big reduction in the Jeffreys posterior predictive probability, and a small reduction in the B-L probability, when reducing the number of successes in the future experiment of size 10 by only one; in the absence of prior information a small reduction would be expected here. Case 5 is a very simple $m = n$ example based on whether $x = n = 4$ is the future or the historical event. We suggest that the close to doubling of the Jeffreys posterior predictive probabilities, favoring the more extreme future event (with no difference between the B-L probabilities), is again undesirable.

B.2 Multinomial model

The multivariate extension of the beta(a, a) prior is a Dirichlet distribution with all parameters equal to a. Based on the uniform distribution that again follows from $a = 1$, Jaynes (2003, p.570) gave the corresponding predictive distribution most elegantly as

$$p(y_1, \ldots, y_c | x_1, \ldots, x_c) = \frac{\binom{y_1 + x_1}{y_1} \cdots \binom{y_c + x_c}{y_c}}{\binom{m + n + c - 1}{m}},$$

where $m = \sum_{i=1}^{c} y_i$ and $n = \sum_{i=1}^{c} x_i$. Again we can write this as a hypergeometric-like distribution with a correction factor (normalisation constant) instead:

$$p(y_1, \ldots, y_c | x_1, \ldots, x_c) = \frac{(n+1)\ldots(n+c-1)}{(m+n+1)\ldots(m+n+c-1)} \frac{\binom{y_1 + x_1}{y_1} \cdots \binom{y_c + x_c}{y_c}}{\binom{m+n}{m}},$$

so that

$$\frac{p(y_1, \ldots, y_c | x_1, \ldots, x_c)}{p(x_1, \ldots, x_c | y_1, \ldots, y_c)} = \frac{(n+1)\ldots(n+c-1)}{(m+1)\ldots(m+c-1)}.$$

As before, we consider it reasonable that when $m = n$, this ratio is 1, which follows from the uniform prior only. For the Dirichlet$(a, \ldots, a)$ prior and $m = n$ this ratio is

$$\frac{p(y_1, \ldots, y_c | x_1, \ldots, x_c)}{p(x_1, \ldots, x_c | y_1, \ldots, y_c)} = \prod_{i=1}^{c} \frac{\Gamma(y_i + a)\Gamma(x_i + 1)}{\Gamma(y_i + 1)\Gamma(x_i + a)}$$

instead. Compare, for example, the extreme outcome $y_1 = n$ with the balanced $x_i \doteq n/c$, which is a generalisation of Case 5 above. It is straightforward to show that for the Jeffreys prior ($a = \frac{1}{2}$) this ratio is $O(n^{\frac{1}{2}(c-1)})$, which thus grows with n as well as c, a clear example of posterior probability being pushed towards the extremes.

C Discussion

This paper has aimed to show that in the case of the binomial model, similar to Bayes' (1763) argument for a uniform prior predictive distribution, the posterior predictive

distribution also implies the uniform prior as the correct representation of prior ignorance. This was achieved by showing desirable symmetry in a hypergeometric-like representation of this distribution, supported by predictive probabilities resulting from the Bayes-Laplace beta$(1,1)$ prior that seem more logical and acceptable than those resulting from the Jeffreys beta$(\frac{1}{2},\frac{1}{2})$ prior, which places undue weight near the extremes.

It may be worthwhile to address the notion that the B-L prior lacks invariance under nonlinear transformations, in contrast with the Jeffreys prior. See, for example, Phillips (1991) and Agresti and Min (2005) who, referring to the latter, stated, "Unlike a uniform prior, it is still the appropriate prior for a one-to-one transformation of the parameter space (e.g., Box and Tiao, 1973, p. 32, 41-42)."

It is easy to see, however, that to adhere to the uniform prior predictive distribution suggested by Bayes (1763), a one-to-one function $\phi = \phi(\theta)$ requires the prior choice $p(\phi) = |d\theta/d\phi|$. Consider, for example, the pdf for the odds $\phi = \theta/(1-\theta)$:

$$p(x|\phi) = \binom{n}{x} \frac{\phi^x}{(1+\phi)^n}, \quad x = 0, \ldots, n, \quad 0 < \phi < \infty.$$

Applying the uniform prior to another function of θ, Fisher (1973, p.16) considered it to contain an "arbitrary element". To apply the uniform prior to just any transformation would seem inadequate indeed, and here $p(\phi) \propto 1$ leads to a non-normalisable posterior for $x = n$ and even $x = n-1$. However, there is nothing arbitrary about the prior $p(\phi) = (1+\phi)^{-2}$, which again leads to $p(x) = (n+1)^{-1}$.

This confirms that invariance is not the unique privilege of the Jeffreys prior, which is based on $p(\phi) = p(\theta)|d\theta/d\phi|$ (Box and Tiao 1973, p.43). We note that Huzurbazar (1976) already gave additional invariance rules that are possible in the presence of sufficient statistics; these rules, when applied to the binomial distribution, lead to either beta$(0,0)$ or beta$(1,1)$ priors.

The invariance implied by Bayes' argument was not made explicit by Edwards (1978) or Stigler (1982). Although it may have been considered to be obvious, it appears that this property is not generally understood. For example, the corresponding prior for the log odds $\eta = \log\{\theta/(1-\theta)\}$ is

$$p(\eta) = \frac{e^\eta}{(1+e^\eta)^2},$$

about which Welsh (1996, p.80) stated: "This is a symmetric density with mode at the origin (corresponding to $\theta = \frac{1}{2}$) which suggests that we believe θ is more likely to be near $\frac{1}{2}$ than 0 or 1. That is, the values of θ are not equally likely on the logit scale and we seem to have prior knowledge about θ." However, applying this argument to the earlier transformation $\phi = \theta/(1-\theta)$ would be to suggest that based on the mode at $\phi = 0$, values of θ near 0 are more likely than values near 1! Clearly all that happens as a result of the odds transformation is that the probability density from $0 < \theta < \frac{1}{2}$ is transferred to $0 < \phi < 1$, and the density from $\frac{1}{2} < \theta < 1$ to $1 < \phi < \infty$, without the introduction of any "prior knowledge about θ" over and above that represented by the uniform prior on θ.

In the context of the relative risk $\rho = \theta_1/\theta_2$, Hashemi et al. (1997) made a similar claim when they stated, "...a uniform prior on the θ_i does not transform to a non-informative prior on the measure of risk. For example, if one assigns a uniform prior to the θ_i, then the density function for ρ has height 0 for $\rho < 0$, $\frac{1}{2}$ in $(0, 1)$ and $1/2\rho^2$ for $\rho > 1$ (that is, the prior for ρ is informative)." Again, this prior is entirely equivalent for $\rho < 1$ and $\rho > 1$, as no information was introduced beyond the (invariant) uniform priors on the θ_i, and here maintains the desirable flat $p(x_1, x_2)$.

The symmetry argument for the binomial's posterior predictive distribution suggested in this paper is similar to Bayes' argument for a uniform prior predictive distribution, again leading to an invariant prior.

References

Agresti, A. and Min, Y. (2005). "Frequentist performance of Bayesian confidence intervals for comparing proportions in 2×2 contingency tables." *Biometrics*, 61: 515–523.

Bayes, T. R. (1763). "An essay towards solving a problem in the doctrine of chances." *Phil. Trans. Roy. Soc. London*, 53: 370–418.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Bernardo, J. M. (1979). "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society, Series B*, 41: 113–147 (with discussion).

— (2005). "Reference analysis." In Dey, D. K. and Rao, C. R. (eds.), *Bayesian thinking: modeling and computation*, 17–90. Amsterdam: Elsevier.

Bernardo, J. M. and Ramon, J. M. (1998). "An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters." *The Statistician*, 47: 101–135.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. New York: Wiley.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. New York: Wiley Classics.

Edwards, A. W. F. (1978). "Commentary on the arguments of Thomas Bayes." *The Scandinavian Journal of Statistics*, 5: 116–118.

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*. New York: Hafner Press.

Geisser, S. (1984). "On prior distributions for binary trials." *The American Statistician*, 38(4): 244–251.

— (1993). *Predictive Inference: An Introduction*. London: Chapman and Hall.

Hashemi, L., Nandram, B., and Goldberg, R. (1997). "Bayesian analysis for a single 2×2 table." *Statistics in Medicine*, 16: 1311–1328.

Huzurbazar, V. S. (1976). *Sufficient Statistics*. New York: Marcel Dekker.

Jaynes, E. T. (2003). *Probability Theory - The Logic of Science*. Cambridge: University Press.

Phillips, P. C. B. (1991). "To criticize the critics: an objective Bayesian analysis of stochastic trends." *Journal of Applied Econometrics*, 6(4): 333–364.

Stigler, S. M. (1982). "Thomas Bayes's Bayesian inference." *Journal of the Royal Statistical Society, Series A*, 145(2): 250–258.

Thatcher, A. R. (1964). "Relationships between Bayesian and confidence limits for prediction." *Journal of the Royal Statistical Society, Series B*, 26: 126–210.

Welsh, A. H. (1996). *Aspects of Statistical Inference*. New York: John Wiley and Sons Ltd.

Acknowledgments

This work was largely undertaken while all authors were members of the Australian Research Council Centre of Excellence in Dynamic Systems and Control, University of Newcastle, Australia. The authors wish to thank an Associate Editor and a Referee for their positive and helpful comments.