UPPSALA
UNIVERSITET

Working Paper 2011:2

Department of Statistics

How Informative is a Non-informative Prior?

Rolf Larsson

Working Paper 2011:2
March 2011
Department of Statistics
Uppsala University
Box 513
SE-751 20 UPPSALA
SWEDEN

Working papers can be downloaded from www.statistics.uu.se

Title: How Informative is a Non-informative Prior?
Author: Rolf Larsson
E-mail: rolf.larsson@statistics.uu.se

How informative is a non-informative prior?

Rolf Larsson

March 25, 2011

Abstract

This paper critizises the notion of non-informative priors. We show that in two well-known discrete cases, binomial and Poisson, the use of priors widely belived to be non-informative, results in posterior intervals that are contained in the corresponding exact frequentist intervals. Moreover, we try to quantify how much information that is carried by the priors by finding out how many observations that need to be added to make the frequentist and Bayesian intervals as close as possible.

Key words: Binomial distribution, Poisson distribution, exact confidence interval, posterior probability interval.

1 Introduction

The purpose of this note is to question the notion of non-informative priors by means of some simple examples: the binomial and Poisson distributions. In these cases, we show that Bayesian posterior probability equal tails intervals are always contained in the corresponding frequentist intervals. Hence, it seems that the 'non-informative' priors carry some information about the unknown parameter. We will try to quantify how informative the priors are by checking how many extra observations that need to be added in order for the frequentist intervals to be as close to the Bayesian intervals as possible.

Related literature is Zhu and Lu (2004), who discuss the binomial example in terms of the Bayes posterior mean estimator for different beta priors. Tibshirani (1989) and Diciccio and Young (2010) investigate which priors and statistical models that are required to get the same posterior and frequentist intervals up to an error that vanishes by increasing the sample size.

The present paper goes on by in turn discussing the Binomial and Poisson distributions, followed by some brief concluding remarks.

2 Binomial distribution

Assume that the random variable X is Binomial(n, p). We will interpret X as the number of successes in n independent trials, each with success probability p.

Denote the observation by x. Let the prior of p be the beta(r, s) distribution, i.e.

$$g(p) \propto p^{r-1}(1-p)^{s-1}$$

where r and s are positive (hyperparameters) and $\propto$ indicates proportionality. The likelihood is

$$L(p;x) = \binom{n}{x} p^x (1-p)^{n-x} \propto p^x (1-p)^{n-x}$$

Hence, the posterior density satisfies

$$\begin{aligned} h(p|x) &\propto p^x (1-p)^{n-x} p^{r-1} (1-p)^{s-1} \\ &= p^{x+r-1} (1-p)^{n-x+s-1} \end{aligned}$$

This means that the posterior distribution of p is beta$(x+r, n-x+s)$. Normed to become a probability distribution, the posterior density is

$$h(p|x) = \frac{\Gamma(n+r+s)}{\Gamma(x+r)\Gamma(n-x+s)} p^{x+r-1}(1-p)^{n-x+s-1}$$

In particular, if $r = s = 1$,

$$\begin{aligned} g(p) &\propto 1, \\ h(p|x) &\propto p^x (1-p)^{n-x}. \end{aligned}$$

i.e. the prior is uniform and the posterior is Beta$(x+1, n-x+1)$. As Zhu and Lu (2004) point out, we may also choose a non informative prior in the Jeffrey's sense by putting $r = s = 1/2$. The motivation behind the latter prior is that it is invariant under reparametrisation.

As for point estimation, the Maximum Likelihood Estimator (MLE) of p is $\widehat{p} = x/n$. For general r and s, the mode of the posterior density is found at

$$p^* = \frac{x+r-1}{n+r+s-2}.$$

Hence, in case $r = s = 1$, we have $p^* = \widehat{p}$. Alternatively, the mean of the posterior distribution is

$$\widetilde{p} = \frac{x+r}{n+r+s}. \tag{1}$$

So as Zhu and Lu (2004) observe, to make $\widetilde{p} = \widehat{p}$, the choice $r = s = 1$ corresponds to adding two observations out of which one is success and one is a failure. Without adding observations, the only way to make $\widetilde{p} = \widehat{p}$ is to put $r = s = 0$, which gives us the limit case of the beta distribution, i.e. a two point distribution with masses $1/2$ on 0 and 1.

The posterior probability interval (a_0, a_1) that covers the parameter with probability $1-\alpha_1-\alpha_2$ is found from solving the equations

$$\frac{\Gamma(n+r+s)}{\Gamma(x+r)\Gamma(n-x+s)}\int_{a_1}^{1} p^{r+x-1}(1-p)^{s+n-x-1}\,dp = \alpha_2, \quad (2)$$

$$\frac{\Gamma(n+r+s)}{\Gamma(x+r)\Gamma(n-x+s)}\int_{0}^{a_0} p^{r+x-1}(1-p)^{s+n-x-1}\,dp = \alpha_1, \quad (3)$$

and specifying $\alpha_1 = \alpha_2 = \alpha/2$, we get an equal tails interval.

The corresponding exact frequentist confidence interval, which is on the form (p_L, p_U), is determined by the equations

$$\sum_{k=x}^{n}\binom{n}{k}p_L^k(1-p_L)^{n-k} = \alpha_1, \quad (4)$$

$$\sum_{k=0}^{x}\binom{n}{k}p_U^k(1-p_U)^{n-k} = \alpha_2. \quad (5)$$

In practice, for given n, x and α (as well as r and s in the Bayesian case), we may solve (2)-(5) numerically to find the corresponding intervals. For example, assume that $n=10$, $x=5$, $\alpha=0.05$ and $r=s=1$. Then, the frequentist equal tails interval is found to be $(0.187, 0.813)$. The corresponding Bayesian equal tails interval is $(0.234, 0.766)$, which is more narrow than, and contained in the frequentist interval. Letting $r=s$ tend to zero, the prior more and more favorises p:s close to zero and one, widening the Bayesian interval. Doing so, the limiting interval turns out to be $(0.212, 0.788)$. However, this interval is contained in the frequentist interval as well. For an asymmetric example, assume instead that $x=8$. Then, the frequentist equal tails interval is $(0.444, 0.975)$. With $r=s=1$, the Bayesian equal tails interval is $(0.482, 0.940)$, and when $r=s$ tends to zero, the interval converges to $(0.518, 0.972)$. Again, these intervals are contained in the frequentist interval.

Proposition 1 shows that this is not a coincidence. Observe that neither equal tails, nor specification of α_1 and α_2 is needed.

Proposition 1 *Taking one observation of Binomial(n,p) random variable, with a beta(r,s) prior for p with $0 < r = s \leq 1$, the Bayesian posterior probability interval for p is always contained in the corresponding frequentist confidence interval.*

Proof. *Assume that $0 \leq r = s \leq 1$. To prove that the lower endpoint of the Bayesian interval exceeds the lower endpoint of the frequentist interval, at first denote the distribution function of a beta(b,c) variable by $F(a;b,c)$ and observe that*

$$\begin{aligned} &\sum_{k=x}^{n}\binom{n}{k}a^k(1-a)^{n-k} \\ = \;& \frac{\Gamma(n+1)}{\Gamma(x)\Gamma(n-x+1)}\int_0^a p^{x-1}(1-p)^{n-x}\,dp = F(a;x,n-x+1). \end{aligned}$$

Hence, the assertion follows from the fact that for all $a \in (0,1)$,

$$F\left(a; x+r, n-x+r\right) \leq F\left(a; x, n-x+1\right),$$

by lemma 3 in the appendix. As for upper endpoints, we similarly have

$$\sum_{k=0}^{x} \binom{n}{k} a^k \left(1-a\right)^{n-k} = 1 - F\left(a; x+1, n-x\right),$$

and the conclusion follows from (cf lemma 3)

$$1 - F\left(a; x+r, n-x+r\right) \leq 1 - F\left(a; x+1, n-x\right),$$

for all $a \in (0,1)$. ■

The observation that adding observations gives the same point estimates in the Bayesian and frequentist cases carries over to confidence intervals. Indeed, it follows from (1) that in case $x = n/2$ and $r = s$, to get the same point estimate we need to add the same number of successes as failures. Focusing on confidence intervals, this number needs to be specified. For example, with $n = 12$, $x = 6$ and $\alpha = 0.05$, we get the frequentist equal tails interval $(0.211, 0.798)$ while $n = 14$ and $x = 7$ yield $(0.230, 0.770)$. The latter is close to the Bayesian equal tails interval with uniform prior ($r = s = 1$) so here, in this sense two extra successes and two extra failures carry the same information as the 'non informative' prior.

With increasing sample size, the prior corresponds to even more extra observations. For example, with $n = 40$, $x = 20$ and $\alpha = 0.05$, the frequentist interval is $(0.338, 0.662)$, while the Bayesian equal tails interval with $r = s = 1$ is $(0.351, 0.649)$. Adding as many successes as failures, the frequentist interval that comes closes is the one with $n = 48$ and $x = 24$, which is $(0.352, 0.648)$, i.e. four extra successes and four extra failures are needed.

It seems that with increased sample size, the number of extra observations corresponding to the prior increases. However, asymptotically, the frequentist and Bayesian intervals coincide (cf Diciccio and Young 2010). In fact, this follows as a corollary to the proposition above, since as $n, x \to \infty$ such that x/n tends to a nonzero constant, while r is held fixed, $F\left(a; x+r, n-x+r\right)$, $F\left(a; x+1, n-x\right)$ and $F\left(a; x, n-x+1\right)$ all behave like $F\left(a; x, n-x\right)$.

3 Poisson distribution

Another example is to let $\mathbf{x} = (x_1, ..., x_n)$ be a sample from the Poisson distribution with (unknown) parameter λ. Then the likelihood is (up to a proportionality constant)

$$L\left(\lambda|\mathbf{x}\right) \propto \lambda^y \exp\left(-n\lambda\right),$$

where $y = \sum_{i=1}^{n} x_i$ is the sufficient statistic. The conjugate distribution is Gamma(θ, β), i.e. the prior fulfills

$$g\left(\lambda\right) \propto \lambda^{\theta-1} \exp\left(-\beta\lambda\right).$$

We see that the choice $\theta = 1$ and $\beta = 0$ gives us an improper uniform distribution, so this would be the natural candidate for a non informative prior. Another choice is $\theta = 1/2$ and $\beta = 0$, giving the Jeffrey's prior (cf Raftery and Akman, 1986). The posterior density is

$$h(\lambda|\mathbf{x}) \propto \lambda^{\theta+y-1} \exp\{-(\beta+n)\lambda\},$$

and so, the posterior distribution is Gamma$(\theta+y, \beta+n)$. In fact,

$$h(\lambda|\mathbf{x}) = \frac{(\beta+n)^{\theta+y}}{\Gamma(\theta+y)} \lambda^{\theta+y-1} \exp\{-(\beta+n)\lambda\}.$$

The MLE of λ is y/n, but the posterior mean is $(\theta+y)/(\beta+n)$, so with $\theta = 1$ and $\beta = 0$, it is $(1+y)/n$. Hence, to make the MLE equal to the posterior mean, we need to increase the observed sum y with one. However, as $\theta \to 0$ with $\beta = 0$, the posterior mean approaches the MLE.

The Bayesian interval (b_0, b_1) is obtained from solving

$$\frac{(\beta+n)^{\theta+y}}{\Gamma(\theta+y)} \int_{b_1}^{\infty} \lambda^{\theta+y-1} \exp\{-(\beta+n)\lambda\} d\lambda = \alpha_2, \quad (6)$$

$$\frac{(\beta+n)^{\theta+y}}{\Gamma(\theta+y)} \int_{0}^{b_0} \lambda^{\theta+y-1} \exp\{-(\beta+n)\lambda\} d\lambda = \alpha_1, \quad (7)$$

while the exact frequentist interval (λ_L, λ_U) solves

$$\sum_{k=y}^{\infty} \frac{(n\lambda_L)^k}{k!} \exp(-n\lambda_L) = \alpha_1, \quad (8)$$

$$\sum_{k=0}^{y} \frac{(n\lambda_U)^k}{k!} \exp(-n\lambda_U) = \alpha_2. \quad (9)$$

In analogy with the beta-binomial example, let us calculate the frequentist equal tails confidence interval and the Bayesian equal tails posterior probability interval when $n = 10$, $y = 5$, $\theta = 1$, $\beta = 0$ and $\alpha = 0.05$. This gives the frequentist interval $(0.162, 1.167)$ while the Bayesian interval is $(0.220, 1.167)$, so again, the Bayesian interval is the most narrow of the two. Moreover, note that the upper endpoints coincide.

As was pointed out above, the posterior mean is larger than the MLE, and to some extent, this seems to explain why the Bayesian interval is a little shifted to the right compared to the frequentist interval. Now, as $\theta \to 0$ with $\beta = 0$, we find that the Bayesian interval approaches $(0.162, 1.024)$, and so, it is still contained in the frequentist interval. Here, it is the lower endpoints that coincide.

In general, we may prove the following proposition.

Proposition 2 *Taking n observations of a Poisson(λ) random variable, with a Gamma(θ, β) prior for λ with $\beta = 0$ and $0 < \theta \leq 1$, the Bayesian posterior probability interval is always contained in the corresponding frequentist confidence*

interval. Moreover, if $\theta = 1$, the upper endpoints coincide, while as $\theta \to 0$, the lower endpoint of the Bayesian interval approaches the lower endpoint of the frequentist interval.

Proof. *Let $0 < \theta \leq 1$ and denote the distribution function of a Gamma$(\theta + y, n)$ variable by $G(b; \theta + y, n)$. Observing that*

$$\frac{n^y}{\Gamma(y)} \int_0^b \lambda^{y-1} \exp(-n\lambda)\, d\lambda = \sum_{k=y}^{\infty} \frac{(nb)^k}{k!} \exp(-nb)$$

and, because of lemma 4 (see the appendix)

$$\begin{aligned} & \frac{n^{\theta+y}}{\Gamma(\theta+y)} \int_0^b \lambda^{\theta+y-1} \exp(-n\lambda)\, d\lambda \\ = \quad & G(b; \theta+y, n) \leq G(b; y, n) = \frac{n^y}{\Gamma(y)} \int_0^b \lambda^{y-1} \exp(-n\lambda)\, d\lambda, \end{aligned}$$

with equality iff $\theta = 0$, we find that if $0 < \theta \leq 1$, then the lower endpoint of the Bayesian interval exceeds the lower endpoint of the frequentist interval. If $\theta = 0$, the endpoints are equal. Similarly,

$$\frac{n^{1+y}}{\Gamma(1+y)} \int_b^{\infty} \lambda^y \exp(-n\lambda)\, d\lambda = \sum_{k=0}^{y} \frac{(nb)^k}{k!} \exp(-nb),$$

and for $0 < \theta \leq 1$,

$$\begin{aligned} & \frac{n^{\theta+y}}{\Gamma(\theta+y)} \int_b^{\infty} \lambda^{\theta+y-1} \exp(-n\lambda)\, d\lambda \\ = \quad & 1 - G(b; \theta+y, n) \leq 1 - G(b; 1+y, n) \\ = \quad & \frac{n^{1+y}}{\Gamma(1+y)} \int_b^{\infty} \lambda^y \exp(-n\lambda)\, d\lambda, \end{aligned}$$

with equality iff $\theta = 1$. Hence, if $0 < \theta < 1$, then the upper endpoint of the Bayesian interval is smaller than the upper endpoint of the frequentist interval, whereas if $\theta = 1$, the endpoints coincide. ■

Now, let us see what happens to the frequentist equal tails interval if we add observations. Keeping the MLE constant at 0.5, the choices $n = 16$, $y = 8$, $\theta = 1$, $\beta = 0$ and $\alpha = 0.05$ give the interval $(0.215, 0.985)$. This is the choice that results in a lower endpoint as close as possible to the corresponding lower endpoint of the Bayesian equal tails interval. However, the upper endpoint is much smaller than for the Bayesian interval, so compared to the binomial case, the situation is less clear-cut here: it is difficult to judge as to how many extra observations that the prior corresponds to. But if the criterion is to get the same lower endpoints, as is reasonable if we focus on one-sided intervals, then in this case the prior corresponds to six extra observations. Increasing the sample size to $n = 40$, if $y = 20$ and $\alpha = 0.05$, then the frequentist interval is $(0.305, 0.772)$,

and if in addition $\theta = 1$, $\beta = 0$, the Bayesian interval is $(0.325, 0.772)$. With the frequentist interval, we need $n = 50$ and $y = 25$ to get as close as possible with the lower endpoint: Then the interval is $(0.324, 0.738)$. So here, in this sense the prior corresponds to adding ten observations.

As in the binomial case, asymptotically the frequentist and Bayesian intervals coincide. This is a corollary of proposition 2, as is seen because as $n, y \to \infty$ such that y/n tends to a nonzero constant and if θ is held fixed, $G(b; \theta + y, n)$ (and in particular $G(b; 1 + y, n)$) behaves like $G(b; y, n)$.

4 Concluding remarks

We have shown that in the binomial and Poisson cases, use of 'non-informative' conjugate priors result in equal tails posterior probability intervals that are contained in the corresponding frequentist confidence intervals. Hence, the 'non-informative' priors seem to carry some information about the unknown parameter. Most probably, this property carries over to other discrete distributions, but this remains to be studied in future research. As for continuous distributions, the same phenomenom does not seem to occur. For example, it is easily seen that the usual frequentist confidence interval for the expectation in the normal distribution, when the variance is fixed, coincides with the corresponding Bayesian interval choosing a flat prior (the normal distribution with 'infinite' variance). Nevertheless, the present paper provides strong arguments against the belief that truly non-informative priors exist in the discrete distribution case.

5 Appendix: Some lemmas

Lemma 3 *Denote the distribution function of a beta(b, c) variable by $F(a; b, c)$. Then, for all $a, r \in (0, 1)$, if $b_1 \leq b_2$ and $c_1 \leq c_2$, we have $F(a; b_2, c_1) \leq F(a; b_1, c_2)$.*

Proof. *Assume $b_1 \leq b_2$ and $c_1 \leq c_2$, and let $U(\gamma, 1)$ be a Gamma$(\gamma, 1)$ random variable. It is well-known that*

$$\frac{U(\gamma_1)}{U(\gamma_1) + U(\gamma_2)}$$

is distributed as beta(γ_1, γ_2). Hence, we need to show that for all $u \in (0, 1)$,

$$P\left\{\frac{U(b_2)}{U(b_2) + U(c_1)} \leq u\right\} \leq P\left\{\frac{U(b_1)}{U(b_1) + U(c_2)} \leq u\right\}$$

But since $U(c_2)$ has the same distribution as $U(c_1) + U(\Delta c)$, where $\Delta c = c_2 - c_1$,

and similarly for $U(b_2)$ *where* $\Delta b = b_2 - b_1$*, this inequality follows from*

$$\begin{aligned}
& \left\{ \frac{U(b_1) + U(\Delta b)}{U(b_1) + U(\Delta b) + U(c_1)} \leq u \right\} \\
= & \{ U(b_1) + U(\Delta b) \leq u [U(b_1) + U(\Delta b) + U(c_1)] \} \\
= & \{ (1-u) [U(b_1) + U(\Delta b)] \leq u U(c_1) \} \\
\subseteq & \{ (1-u) [U(b_1)] \leq u [U(c_1) + U(\Delta c)] \} \\
= & \{ U(b_1) \leq u [U(b_1) + U(c_1) + U(\Delta c)] \} \\
= & \left\{ \frac{U(b_1)}{U(b_1) + U(c_1) + U(\Delta c)} \leq u \right\}.
\end{aligned}$$

■

Lemma 4 *Denote the distribution function of a Gamma*(θ, β) *by* $G(b; \theta, \beta)$*, where* $\theta, \beta > 0$*. Then, for* $\theta_1 \leq \theta_2$*, we have* $G(b; \theta_2, \beta) \leq G(b; \theta_1, \beta)$*.*
Proof. *Assume* $\theta_1 \leq \theta_2$*, and let* U_1 *be distributed as Gamma*(θ_1, β) *and* U_2 *be distributed as Gamma*$(\theta_2 - \theta_1, \beta)$ *independently of* U_1*. Then,* $U_1 + U_2$ *is Gamma*(θ_2, β)*, and we have for all* $b > 0$ *that*

$$G(b; \theta_2, \beta) = P(U_1 + U_2 \leq b) \leq P(U_1 \leq b) = G(b; \theta_1, \beta).$$

■

6 References

Diciccio, T. and Young, G.A. (2010), "Objective Bayes and conditional inference in exponential families," *Biometrika,* 97, 497-504.

Raftery, A.E. and Akman, V.E. (1986), "Bayesian Analysis of a Poisson Process with a Change-Point," *Biometrika,* 73, 85-89.

Tibshirani, R. (1989), "Noninformative priors for one parameter of many," *Biometrika,* 76, 604-608.

Zhu, M. and Lu, A.Y. (2004), "The Counter -intuitive Non-informative Prior for the Bernoulli Family," *Journal of Statistics Education,* 12, 1-9.