Electronic Journal of Statistics
Vol. 5 (2011) 1450–1470
ISSN: 1935-7524
DOI: 10.1214/11-EJS648

Neutral noninformative and informative conjugate beta and gamma prior distributions

Jouni Kerman

Novartis Pharma AG
CH–4002 Basel, Switzerland

Abstract: The conjugate binomial and Poisson models are commonly used for estimating proportions or rates. However, it is not well known that the conventional noninformative conjugate priors tend to shrink the posterior quantiles toward the boundary or toward the middle of the parameter space, making them thus appear excessively informative. The shrinkage is always largest when the number of observed events is small. This behavior persists for all sample sizes and exposures. The effect of the prior is therefore most conspicuous and potentially controversial when analyzing rare events. As alternative default conjugate priors, I introduce Beta(1/3, 1/3) and Gamma(1/3, 0), which I call 'neutral' priors because they lead to posterior distributions with approximately 50 per cent probability that the true value is either smaller or larger than the maximum likelihood estimate. This holds for all sample sizes and exposures as long as the point estimate is not at the boundary of the parameter space. I also discuss the construction of informative prior distributions. Under the suggested formulation, the posterior median coincides approximately with the weighted average of the prior median and the sample mean, yielding priors that perform more intuitively than those obtained by matching moments and quantiles.

Keywords and phrases: Prior distributions, noninformative distributions, Bayesian inference, conjugate analysis, beta distribution, gamma distribution.

Received September 2010.

1. Overview

1.1. Introduction

The focus of this paper is on the choice of a default (noninformative) or an informative prior distribution for the two well-known conjugate models: the binomial-beta model $y \sim \text{Binomial}(n, \theta)$ with a prior distribution $\theta \sim \text{Beta}(a, b)$, and the Poisson-gamma model $y \sim \text{Poisson}(\lambda X)$ with the $\lambda \sim \text{Gamma}(c, d)$ conjugate prior, where X is the given exposure. Here, the beta distribution is parameterized as being proportional to $\theta^{a-1}(1-\theta)^{b-1}$ and the gamma distribution proportional to $\lambda^{c-1}\exp(-d\lambda)$. For the binomial-beta model, the posterior for the unknown proportion or event rate θ is $\text{Beta}(a+y, b+n-y)$, so that a and b are often interpreted as the prior number of successes and failures, respectively.

The posterior of the unknown rate λ in the Poisson-gamma model is again gamma-distributed, $\text{Gamma}(c+y, d+X)$, where c is often interpreted as the prior number of observations within an exposure of d. Here, a "noninformative" prior does not necessarily refer to a flat or a uniform prior, but to a distribution that lets the likelihood play a major role in forming the posterior distribution.

Conjugate models are computationally convenient, since they yield posteriors in the same distributional family as the priors and are also conceptually straightforward as they allow a direct interpretation of the distribution parameters as functions of the data. For these two models, the posterior mean can be computed directly as the weighted average of the prior mean and the sample mean, with the weight of the prior mean being the proportion of the prior sample size (or exposure) to the combined sample size (or exposure). Hence, the effect of the prior is usually assessed in terms of the shrinkage of the sample mean: a prior with little information is associated with a small weight on the prior mean and a large weight on the sample mean.

Looking only at how the distribution mean changes does not necessarily give a full picture of the behavior of the prior. Judging the degree of informativeness by the shape or prior variance is not always helpful either. In the case of estimating rare events where y is small, say 0, 1, or 2, the distribution of the posterior probability mass relative to the scale (implied by sample size n or exposure X) will be affected by the choice of the prior. In particular, assuming the uniform beta prior for the binomial model, as much as 74% of posterior mass may be above the observed sample mean. Under a different prior, most of the posterior mass may lie below the sample mean instead. This tendency persists even if the sample size n (or exposure X) is large: the shrinkage is due to the prior and the absolute number of successes observed, y. Exactly the same phenomenon can be witnessed in the gamma-Poisson model.

This brings us to the question of the choice of a noninformative prior distribution, and specifically the choice of a suitable default prior. Such a prior should be applicable to all possible outcomes, sample sizes, and exposures. In practice, even though prior distributions should reflect the current state of scientific knowledge, default priors can be considered as 'reference priors,' being "merely technical devices to facilitate the derivation of reference posteriors" [7]. The quantiles and predictive probabilities of such priors are not necessarily of interest a priori.

When analyzing data from testing a novel drug, for example, a default noninformative distribution is invariably used. In these situations, statisticians tend to choose 'off-the-shelf' priors and trust that since they are considered generally to be noninformative, they should not introduce any extraneous information that is not conveyed by the data.

It is then important, especially to applied statisticians and non-statisticians (such as clinicians) involved in the design and interpreting the results of the trial, that a noninformative default distribution performs according to their expectations. A default distribution must produce intuitive posterior inferences, otherwise the choice of prior cannot be justified as this would mean that the model conflicts with the implicit prior knowledge of the investigators. For example, if after observing one failure out of one Bernoulli trial, concluding that $\Pr(\theta < \epsilon | y = 0, n = 1) \approx 1$ for some arbitrarily small given ϵ would certainly be considered to be excessively informative, but this is however possible by adopting a prior $\mathrm{Beta}(a, a)$ with a sufficiently small $a \approx 0$. Even such extreme distributions are often presented as noninformative in the literature, although they have been also severely criticized [20].

The role of the noninformative gamma and beta distributions is, essentially, to regularize inferences by assigning little weight to extreme assumptions, avoiding extreme posterior statements that we find hard to believe *a priori*. The effect of these priors can be characterized as 'default' information, being most conspicuous in the case of $y = 0$ (or $y = n$), as the posterior quantiles, relative to the scale (n or X), depend only on the information provided by the prior. Essentially, they are all 'weakly informative' distributions as characterized by Gelman et al. [10]; calling the priors 'noninformative' is thus more of a convention. The choice of a prior therefore always involves a choice of degree of informativeness, while being comfortable with the resulting posterior inferences for each possible outcome. Therefore, one must decide how much 'default information' about the background rate in the model is appropriate. As it will be shown in this paper, this default information has a certain effect on the posterior inferences even if the underlying prior is commonly considered to be noninformative.

1.2. Evaluating the posterior performance of a prior

To measure the effect of the prior on posterior inferences, an obvious way is to compare the posterior distribution of, say θ, with the maximum likelihood estimate $\hat{\theta}$ (here, the sample mean) by computing the tail probability $\Pr(\theta > \hat{\theta} | y)$. This probability is then compared to $1/2$, which would follow in the case of a symmetric likelihood and a flat prior. For example, when the sample mean is transformed into the logistic scale (assuming it is not 0 or 1) and the rate is modeled using a Normal distribution with the usual uniform prior over the real line, the tail probability will be exactly $1/2$. Still, it is natural and convenient to work on the original scale using conjugate models if no covariate information is available or to be used; the problem of choosing a suitable default prior distribution will not go away by transforming the problem to another scale. It will be shown that in the conjugate models, the posterior distribution will not be necessarily evenly distributed around the MLE, producing tail probabilities that may differ considerably from $1/2$. This holds especially when y is small, even for large sample sizes or exposures.

In early-phase clinical trial designs, go/no-go criteria are often formulated using Bayesian posterior tail area probabilities [16]. For example, one may claim that a drug is promising if $\Pr(\theta > \tau | y)$ is large enough for some predetermined fixed threshold τ. Hence, it is crucial to understand how the choice of prior can affect the quantiles. If $\Pr(\theta > \hat{\theta} | y)$ is large or small under a supposedly noninformative prior, it tends to translate to a correspondingly large or small tail probability involving any given threshold τ as well.

1.3. Posterior neutrality

Taking the tail probability $\Pr(\theta > \hat{\theta}|y)$ as a measure of the effect of the prior, it must then be asked how this would be affected under a noninformative prior. As usual, it is important to have a distribution that will translate to a relatively large posterior variance, representing the effect of a prior with large uncertainty. Still, this is not enough, as we need to consider the location of the posterior mass as well, the natural point of reference being the maximum likelihood estimate. It is intuitive that under a noninformative prior, $\Pr(\theta > \hat{\theta}|y)$ should not be too close to 0 nor too close to 1. If the tail probability were consistently near 1/2, the prior could be characterized as 'neutral' as then it would not appear to favor neither the case $\theta > \hat{\theta}$ nor the case $\theta < \hat{\theta}$.

The priors Beta$(1/3, 1/3)$ and Gamma$(1/3, 0)$ have the special property that the posterior probability $\Pr(\theta > \hat{\theta}|y)$ (or $\Pr(\lambda > \hat{\lambda}|y)$ in the Poisson model) is approximately 0.5 for all possible outcomes, sample sizes, and exposures, as long as the sample mean is not at the boundary of the sample size ($y = 0$ or $y = n$). This implies that the maximum likelihood estimate (sample mean) is then approximately at the posterior median. Further, the beta prior has a prior variance that is larger than the common Jeffreys and Uniform priors.

A neutral beta or gamma prior distribution has the property of yielding a posterior median as an approximate weighted average of the prior median and the sample mean. The approximation is quite accurate as long as the actual number of observed events (successes) y is at least 1 (and at most $n-1$ in the case of the binomial model). This property allows us to construct informative priors by 'matching the median,' producing prior distributions that behave consistently and in an intuitive manner.

2. Noninformative priors

2.1. Noninformative beta priors

In practice, any distribution of the form Beta(a, a) is considered as a potential choice for a noninformative prior for the binomial model as long as $0 \leq a \leq 1$. The two shape parameters of the distribution match the successes and failures of the binomial likelihood, and therefore their sum $2a$ is usually characterized as the prior sample size. Symmetry (equality of both shape parameters) is a natural requirement for a default prior distribution: such a prior must apply to modeling both successes and failures without yielding conflicting inferences. However, there is no universal consensus about which Beta(a, a) distribution one should choose by default. The three most popular choices are the Haldane prior Beta$(0, 0)$ the Jeffreys prior Beta$(1/2, 1/2)$ and the uniform (Bayes-Laplace) prior Beta$(1, 1)$.

The uniform Beta$(1, 1)$ prior has been almost invariably introduced in textbooks on Bayesian statistics in elementary examples [e.g., 9, 19]. This distribution has the most appealing form (on the probability scale) and, apparently, represents lack of prior information by assigning equal weight to all possible parameter values.

The Jeffreys prior $\mathrm{Beta}(1/2, 1/2)$, obtained by applying the Jeffreys's rule [13] on the binomial likelihood is another well-respected candidate for a reference prior. Jeffreys's rule has been advocated especially by Bernardo [6] and the proponents of objective Bayesianism [5], the $\mathrm{Beta}(1/2, 1/2)$ prior for the binomial likelihood being one of the recommended reference priors.

The improper Haldane prior [11], $\mathrm{Beta}(0, 0)$, represents the limit as the implied sample size $2a$ tends to zero. Moreover, only the Haldane prior yields posteriors with the posterior mean $(a+y)/(2a+n)$ exactly at the sample mean y/n. This is an attractive property, but the prior yields improper posterior distributions for the extreme outcomes $y = 0$ and $y = n$. Also, approximate Haldane priors of the form $\mathrm{Beta}(\epsilon, \epsilon)$ are unintuitive as they put almost all posterior mass in either of the endpoints of the parameter interval. Priors with either shape parameter less than 1, especially those close to zero, have been criticized by Tuyl, Gerlach and Mengersen [20], showing by example how posterior distributions seem to be affected too much by the information in the prior. However, as we shall see, also a beta prior with $a \approx 1$ may appear to be excessively informative.

Figure 1 shows a plot of the posterior tail probabilities $\Pr(\theta > y/n|y)$ for the three conventional priors and the proposed Neutral prior for the case of $n = 40$. Looking at the outcomes for $y = 1, \ldots, n-1$, the difference of the tail probability and $1/2$ is always largest at the points $y = 1$ or $y = n-1$, and the smallest (zero) at $y = n/2$. This holds for any $n \geq 3$.

Since the effect of the prior in terms of mean shrinkage diminishes as the sample size n increases, one may mistakenly also assume that the posterior will be approximately neutral for large n. However, this will be true only for outcomes y close to $n/2$. For any given fixed y, the tail probability $\Pr(\theta > y/n|y)$ approaches a limit, $\Pr(\gamma > y|y)$, where $\gamma \sim \mathrm{Gamma}(a+y, 1)$, which depends on $a+y$, so the tail probability will not necessarily be $1/2$. For example, assuming $y = 1$ and $a = 1$, $\Pr(\theta > y/n|y)$ increases from 0.593 to 0.736 as n increases from 3 to ∞. The pattern shown in Figure 1 holds roughly for other sample sizes as well: for a given y, the tail probability tends to a constant as n increases. The posterior distributions for a given y tend to be centered away from y/n unless a is close to $1/3$.

It is illustrative to look at an example of how posterior quantiles vary by the choice of prior. Figure 2 shows how the prior influences the posterior quantiles in the case of $y = 1$ for large n. The middle posterior 95% and 50% intervals (with the endpoints at the 0.025, 0.25, 0.75, and 0.975 quantiles) are shown along with the medians (as dots). Depending on the choice of prior, all quantiles are either pulled toward the prior median or away from it. For larger y, the relative distance between the medians is smaller; in the special case of $y = n/2$ the median coincides exactly at y/n. The Neutral prior has the property of producing posteriors with y/n overall the closest to the median for any given y and n as long as $y/n \in (0, 1)$. Figure 3 shows the corresponding posterior densities, illustrating the skewness of the posterior distributions.

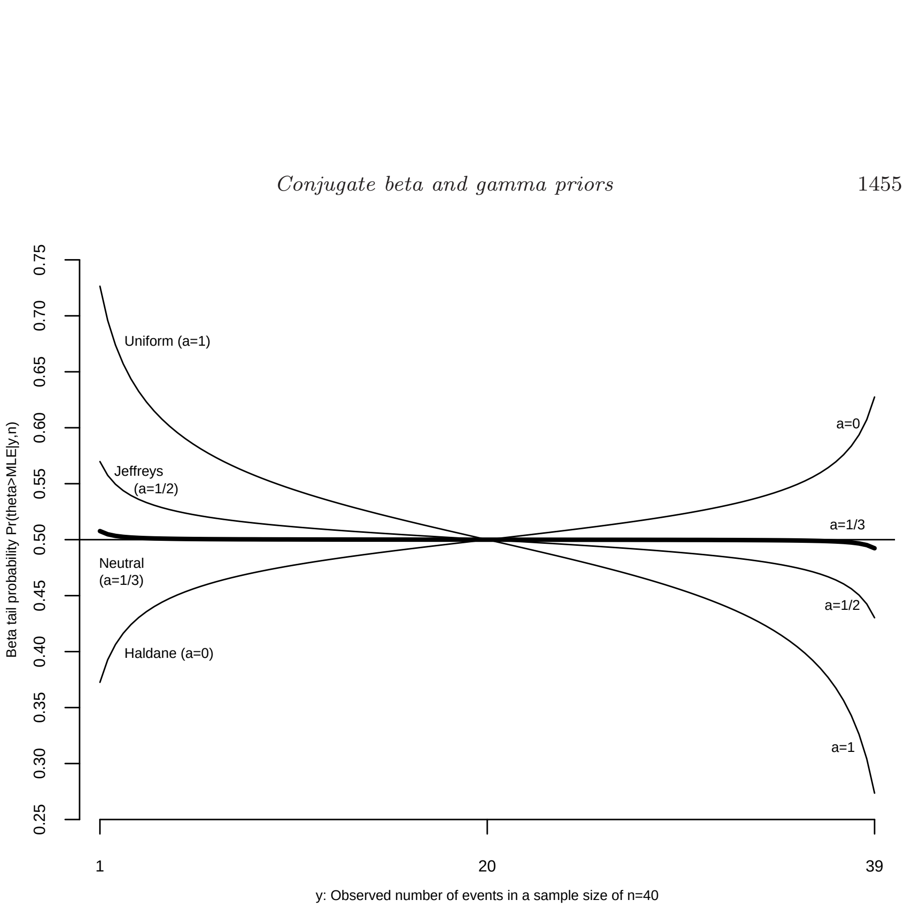

FIG 1. *Posterior tail probabilities* $\Pr(\theta > y/n|y)$ *for outcomes* $y = 1, \dots, n-1$*, assuming a prior* $\theta \sim$ *Beta*(a,a) *and a sample size* $n = 40$*. The neutral prior Beta*$(1/3, 1/3)$ *has the unique property of centering the posterior distribution almost exactly at the sample mean, while other symmetric beta priors with the shape parameter* $a \leq 1$ *tend to shift the posterior mass either to the left or to the right of the point estimate, depending on the outcome.*

2.2. Noninformative gamma priors

Many supposedly noninformative gamma conjugate priors are used in practice, for example the improper but scale-free prior $\text{Gamma}(1/2, 0)$ obtained by the Jeffreys's rule, $\text{Gamma}(\epsilon, \epsilon)$ (with $\epsilon \approx 0$), and even $\text{Gamma}(\epsilon, 1)$, the rationale being that as $\epsilon \to 0$, the distribution appears flatter and flatter over the positive real axis. However, at the same time the probability mass concentrates near zero. For example, assume a $\text{Gamma}(0.01, 0.01)$ prior for the Poisson conjugate model. After observing one outcome during an exposure of 40 units, the posterior is $\text{Gamma}(1.01, 40.01)$. The posterior median (0.0176) is located to the left of the sample mean 0.025 with a tail probability $\Pr(\lambda > 1/40) = 0.37$. Under the Jeffreys prior $\text{Gamma}(1/2, 0)$, the median is located to the right of the mean, at 0.03, with a tail probability of 0.572.

Figure 4 shows a plot of posterior tail probabilities $\Pr(\lambda > y|y)$ for several prior distributions assuming a fixed exposure $X = 1$ (as the units of exposure can be chosen arbitrarily) Under priors $\text{Gamma}(c, d)$, only those with d exactly zero (the scale-free improper priors) produce posteriors whose tail probabilities converge eventually to $1/2$ as $y \to \infty$.

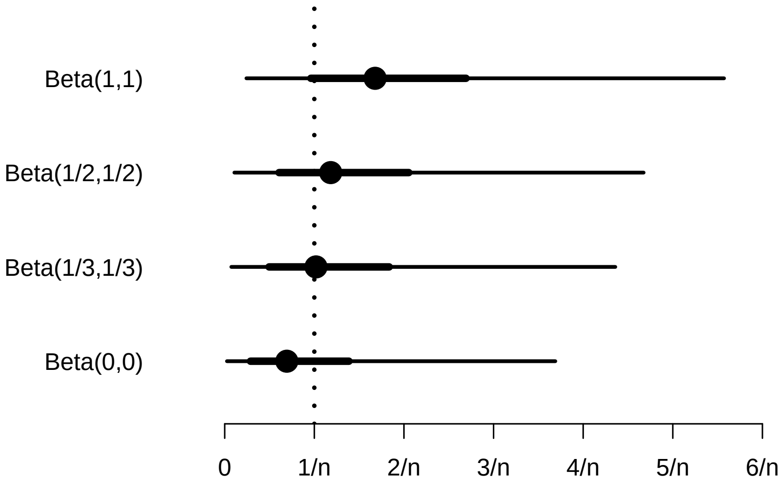

Fig 2. *Middle posterior 95% intervals (thin lines), middle 50% intervals (thick lines), and medians (dots) for four conjugate beta models assuming different prior distributions for the case of one single observed success ($y = 1$) for large n. This graph shows how the effect of the chosen default prior affects the distribution of the posterior mass relative to the sample mean (maximum likelihood estimate) $1/n$, which is highlighted as a vertical dotted line. As n increases, the posterior distribution $Beta(a + y, a + n - y)$ tends to a $Gamma(a + y, n)$ distribution, so this graph equivalently illustrates the corresponding gamma posteriors under four different priors $Gamma(a, 0)$ when the exposure equals some given number $n > 0$.*

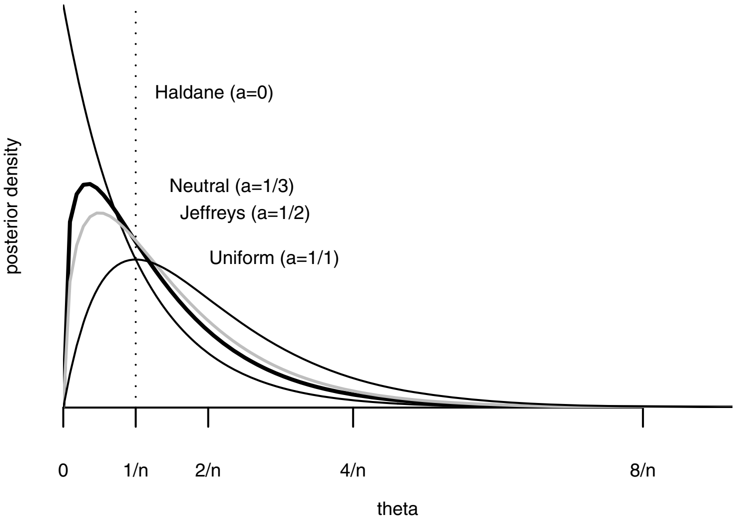

Fig 3. *Beta posterior densities in the case of $y = 1$ and large n, based on the four beta conjugate priors, illustrating the skewness of the resulting posterior distributions. The maximum likelihood estimate $1/n$ is highlighted by a vertical dotted line.*

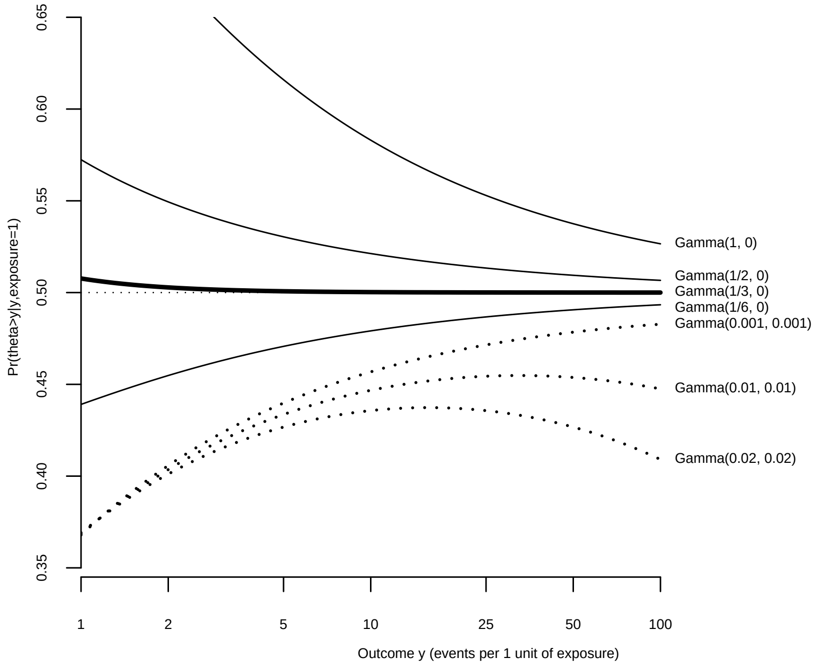

FIG 4. *Posterior tail probabilities* $\Pr(\lambda > y|y)$ *(with* $X = 1$*) under different gamma prior distributions. Priors of the form Gamma*(ϵ, ϵ) *(dotted lines) produce posteriors that are considerably skewed in relation to the MLE (sample mean). The tail probabilities under scale-free priors (with rate parameter equal to zero) converge eventually to* $1/2$*, but those with even a small but nonzero prior exposure (dotted lines) will eventually converge to zero, showing the influence of the prior exposure.*

The tail probability $\Pr(\lambda > y|y)$ under a prior with even a small but nonzero prior inverse scale $d > 0$ will eventually tend to zero. The probability first approaches 0.5, then converges to zero, although slowly. This is because the probability $\Pr(\lambda > y/X|y)$ depends on d/X: $\Pr(\lambda > y/X) = \Pr(\lambda^* > y(1 + d/X))$ where $\lambda^* \sim \text{Gamma}(c + y, 1)$. Thus if $d > 0$, then as the precision of the parameter increases, the tail probability will tend to zero as λ^* tends to y. Using a smaller scale, and consequently large values of X, would result in suppressing the effect of a nonzero d at first, but for large y the tail probability will eventually always tend to zero. For a default conjugate prior, the choice of $d > 0$ cannot be justified. If specifying an improper prior is not possible, d must be chosen such that yd/X is close to zero for all practically possible values of (y, X). All proper prior gamma distributions involve the choice of the inverse scale parameter d, and therefore also the choice of the unit of the scale.

2.3. A neutral beta prior

According to the approximation introduced by Kerman [15], the median of the Beta(a, b) distribution is approximately,

$$\frac{a-1/3}{a+b-2/3},$$

which approaches asymptotically the median when both shape parameters increase. A posterior distribution of the form $\mathrm{Beta}(a+y, a+n-y)$ has therefore the median approximately equal to,

$$\frac{y+(a-1/3)}{n+2(a-1/3)};$$

hence choosing $a = 1/3$ yields the median approximately at the sample mean y/n. This prior yields posterior medians close to y/n for all possible sample sizes n and all $y = 1, \ldots, n-1$, with the posterior tail probabilities,

$$\Pr(\theta > y/n|y, n) \in [1/2, 0.508)$$

for all $n \geq 2, y/n \in (0, 1/2]$, and, by symmetry,

$$\Pr(\theta \leq y/n|y, n) \in (0.492, 1/2)$$

for $y/n \in (1/2, 1)$. As y increases while the proportion y/n stays constant, the tail probability tends asymptotically to 1/2 from above when $y/n < 1/2$. Approximations such as $a \approx 0.33$ yield similar tail probabilities, but convergence to 1/2 is slower, and the tail probability is not necessarily greater than 1/2 whenever $y/n < 1/2$. When $a = 1/3$, the absolute distance between the median and y/n will also be closer to zero for large n.

In the case of a trial with 1 event out of 40 observations, the posterior based on the $\mathrm{Beta}(1/3, 1/3)$ distribution is $\mathrm{Beta}(1/3 + 1, 1/3 + 40 - 1)$, with the tail probability $\Pr(\theta > 1/40) = 0.508$, compared to 0.373, 0.570, and 0.726 for the Haldane, Jeffreys and uniform priors, respectively. The tail probability approaches 1/2 from above as y increases to $n/2$; by symmetry as $y \to n$, the tail probability correspondingly decreases. Moreover, the tail probability approaches 1/2 very fast: if $5 \leq y \leq n-1$, $\Pr(\theta > y/n) \in (0.499, 0.501)$ for all n. This distribution also has larger variance than than those of the Jeffreys and the uniform beta distributions.

2.4. A neutral gamma prior

Approximations to the gamma median when the shape parameter is continuous were investigated by Berg and Pedersen [4]. For a $\mathrm{Gamma}(c, d)$ distribution, the median is approximately,

$$\frac{c-1/3}{d}.$$

This form implies that a posterior distribution of the form $\mathrm{Gamma}(c+y, d+X)$ (with $y \geq 1$) has a median approximately equal to,

$$\frac{y+(c-1/3)}{d+X};$$

choosing $c = 1/3$ and $d = 0$ yields the best approximation for y/X.

Using this approximation, the relative error of the tail probability $\Pr(\lambda < y/X|y, X)$ is at most 4% for $y \geq 1$. The error decreases rapidly; for $y \geq 2$ the error is less than 1%.

Figure 4 shows that $\mathrm{Gamma}(1/3, 0)$ produces posteriors that are almost exactly centered at y/X for even small y. Although the prior is improper, the exposure X will be necessarily positive, so the posterior will also be necessarily proper.

Under $\lambda \sim \mathrm{Gamma}(1/3, 0)$,

$$\Pr(\lambda > y/X|y, X) \in (1/2, 0.508)$$

for all $y \geq 1, X > 0$. As y increases, the gamma posterior tail probability approaches 1/2 asymptotically from above.

3. Informative priors

3.1. Informative beta priors

Suppose that we have observed a proportion $p = y/m$ in a trial of size m. For the analysis, a symmetric default prior $\mathrm{Beta}(a, a)$ is used, with some conventional choice of $a > 0$. However, when constructing a beta prior from past data (y, m), in practice this is often done by matching the point estimate to the mean of the distribution [see e.g., 17], that is, constructing a prior with an exact mean $p = y/m$, but this is possible only if we assume that the 'prior of the prior' is the improper Haldane prior $\mathrm{Beta}(0, 0)$:

$$\theta \sim \mathrm{Beta}(a + mp, \quad a + m(1 - p)),$$

with $a = 0$. For an analysis assuming no specific prior information, a shape parameter $a > 0$ is practically always used, but when a prior based on data from a previous trial is derived by matching the mean, we implicitly set $a = 0$ (unless we actually see no events). In other words, the choice of $a > 0$ is often done mainly to avoid an improper posterior in the case $y = 0$ or $y = n$, but if the trial yields $p \in (0, 1)$, the implied noninformative prior is switched to the Haldane prior with $a = 0$. This is inconsistent and looks suspicious to a non-Bayesian; seasoned Bayesians seem to be used to it, trusting that a prior with an implied small prior sample size should not matter in the end. However, even a relatively large sample size does not guarantee that the posterior is close to neutral. Tuyl, Gerlach and Mengersen [20] show by example how prior distributions where one or both shape parameters are close to zero may also dominate the likelihood with a larger sample size than implied in the prior.

For consistency, all beta priors should be based on a noninformative, default prior that provides the background default information for the model. Under the Neutral prior ($a = 1/3$), priors would thus always take the form,

$$\theta \sim \mathrm{Beta}(1/3 + mp, \quad 1/3 + m(1 - p)), \tag{1}$$

where m would necessarily be nonnegative, with the case $m = 0$ yielding the neutral, noninformative prior. For an arbitrary Beta(a, b), this would imply that $m = a + b - 2/3$ and $p = (a - 1/3)/m$. If $m = a + b - 2/3 = 0$, the distributional form (1) does not exist, unless $a = b = 1/3$.

Under this formulation, an arbitrary Beta(a, b) distribution can be interpreted to have an implied sample size of $m = a + b - 2/3$. The information provided by the default noninformative prior is zero by convention and not counted as part of the implied prior sample size. Similarly, if one wishes to use the Uniform prior as the default, the implied sample size would be $m = a+b-2$, and under the Haldane prior, this would be $m = a + b$.

When the information (m, p) is carried over to the next trial in the form of (1), the resulting posterior distribution will have the median approximately at the weighted average of the sample mean and p. If (n, q) are the newly observed data, the posterior will be,

$$\text{Beta}(1/3 + r(m + n), 1/3 + (1 - r)(m + n)),$$

where $r = pw + q(1 - w)$ and $w = m/(m + n)$. In other words,

$$\text{posterior median} \approx \frac{m}{m + n} p + \frac{n}{m + n} \text{sample mean.}$$

In particular, if the observed proportion q coincides with that in the prior trial, the posterior median will again be approximately equal to $r = q = p$, but with a larger implied sample size $m + n$, reinforcing the posterior belief that the distribution is centered at p.

This approximation has a relative error of at most 4% when the posterior shape parameters are both at least 1; when they are at least 4/3 the error is already less than 2%. The error approaches zero fast as the shape parameters increase. In these cases p also has the interpretation as the approximate prior median. Also, under the Neutral prior, the absolute distance between the MLE and the posterior median approaches zero very fast compared to posteriors based on other priors.

The distributional form (1) gives us insight to explaining the behavior of the conventional noninformative priors. The uniform Beta$(1, 1)$ prior implies $p = 0.5$ and $m = 4/3$. The posterior median under the uniform prior is therefore approximately equal to the weighted average of the sample mean and $p = 0.5$ with a substantial weight $m = 4/3$, hence the shift toward the middle of the parameter space. The Jeffreys prior is similarly obtained with $p = 0.5$ but with a smaller implied sample size $m = 1/3$. The Haldane prior can then be constructed by using $p = 0.5$ and a negative sample size of $m = -2/3$; the negative sample size implies that the posterior median is pulled away from the prior median, toward either boundary.

When constructing an informative prior to use in an actual trial, one should never take the prior data at face value. Rather, one should down-weight the amount of information obtained in the previous trial by choosing an m smaller than the total prior sample size, thus including some between-trial variation that should increase our uncertainty. As the prior sample size m approaches zero, the prior median will shift toward 0.5 as the distribution will be then closer to the symmetric Beta$(1/3, 1/3)$ prior. The prior with $m = 0$ would then be equivalent to no prior information. Of course, m must be chosen considering the size n of the future trial, since the prior weight $w = m/(m + n)$ depends on both m and n. As Box and Tiao [8] noted, "the statement 'knowing a little a priori' can only have meaning relative to the information provided by an experiment."

3.2. Informative gamma priors

The general principles presented above also apply to the construction of informative gamma priors. The information contained in the number of observations y and the exposure X can be down-weighted by a factor $k \in [0, 1]$, using the distributional form,

$$\lambda \sim \text{Gamma}(1/3 + ky, \quad kX). \tag{2}$$

$k = 0$ yields the noninformative prior Gamma$(1/3, 0)$ while $k = 1$ implies full weight. The implied prior rate is always $r = y/X$. If a rate $r' = y'/X'$ is observed, the posterior will be approximately centered at the weighted average of the observed rates $rw+r'(1-w)$, where $w = kX/(kX+X')$. The elicitation of a gamma prior is most naturally done by eliciting the number of events y during an exposure of d units. The scale of the exposure can be chosen arbitrarily.

To avoid overly high prior precision and to make the prior robust enough for the purposes of the future trial, the down-weighting factor k should be used appropriately, as the degree of robustness of the prior depends on the prior weight $w = kX/(kX + X')$, which is a function of the exposure X' to be witnessed in the future trial. The only down-weighted gamma distribution that is robust enough against all possible trials is the noninformative distribution Gamma$(1/3, 0)$ itself.

4. Implications of the choice of priors

4.1. The case of zero events

In the case of $y = 0$ (also, $y = n$ for the binomial model), the posterior is extremely influenced by the choice of prior: one can say that the likelihood provides no evidence that the rate differs from zero, so the prior must provide all the evidence to the contrary. This can be seen by letting $a \to 0$: all quantiles of the posterior distribution Beta$(a + 0, a + n)$ (and of Gamma$(a + 0, 0 + X)$) collapse to zero. These posterior inferences are due solely to the influence of the prior: from the Bayesian point of view, the posterior quantiles themselves are essentially arbitrary unless the underlying prior is considered informative, at least to some degree. Perhaps the only reasonably 'objective' assessment of the case of zero events that can be done is that $\Pr(\theta < 1/n|y)$ should be relatively high. Intuitively, this probability should be closer to 1 than to 0.5, which is the reference tail probability for the case $y = 1$.

Still, there have been attempts to extract some information from the likelihood only: the so-called 'rule of three' [12, 14], gives a frequentist upper limit of the 95% confidence interval of the binomial proportion, which is approximately $3/n$. Of course, this does not reflect the true state of nature, which would be attempted when using a Bayesian model. For someone assessing the true state of nature of θ, the rule of three is but a convention agreed beforehand, and not something that can be inferred from the data. The Bayesian equivalent of the rule of three [14] is derived from the 95th percentile of the beta posterior distribution assuming the uniform prior ($a = 1$) and no events out of n. This is exactly $1 - 0.05^{1/(n+1)}$, being bounded from above by $-\log(0.05)/(n+1) \approx 3/(n+1)$. Similarly, assuming another prior one may obtain different rules. For $a = 1/3$ we can, for example, obtain a crude 'rule of three halves': the 95th percentile is bounded from above by $1.5/(n+0.5)$; a more exact bound would be $u/(n+u/2)$ where $u = 1.47328$ is the 95th percentile of the $\mathrm{Gamma}(1/3, 1)$ distribution.

Under the uniform beta prior distribution, the 95% posterior quantile is the largest among the noninformative priors $\mathrm{Beta}(a, a)$ with $a \le 1$; consequently the prior has been characterized as "conservative" [20]. As discussed, the apparent conservativeness is due to the weight at $\theta = 0.5$ that pulls the median and other quantiles toward the middle of the parameter space. Thus any symmetric prior with $a > 1$ yields an even larger 95% quantile (for a given n). Similarly, the other quantiles are affected as well.

To be able to make posterior statements with confidence, an appropriate informative prior must be used. Still, one must be comfortable with the default inferences provided by the default prior. The posterior median under the Neutral prior is approximately $0.1/n$, while under the uniform prior the median is $0.7/n$. If for a large n there are no events, the Neutral prior gives $\Pr(\theta < 1/n | y = 0) = 0.904$, while the uniform prior gives a much lower probability 0.632.

The neutral gamma prior yields similar probabilities, as the limit of $\Pr(\theta < y/n|y)$ is a gamma tail probability: for large n, the distribution $\mathrm{Beta}(a, a+n)$ is well approximated by a $\mathrm{Gamma}(a, n)$. For $a = 1$, the $\mathrm{Gamma}(a+0, 0+X)$ posterior distribution has a median at $\log(2)/X = 0.693/X$ and the 95% percentile at $-\log(0.05)/X \approx 3/X$, a 'rule of three'. For the neutral $\mathrm{Gamma}(1/3, 0)$ prior, one can also obtain a 'rule of three halves,' as the upper bound of the 95th percentile is approximately $1.47328/X$.

Figure 5 illustrates the effect of the default prior in case of $y = 0$ in the limit when $n \to \infty$. The interval's right endpoint is at the posterior 95% quantile, with the dot marking the median.

4.2. Frequentist coverage properties

Due to the discrete nature of the data in the binomial and Poisson models, the conditional frequentist coverage probability is not constant over the possible hypothetical fixed values of θ or λ: plots show a familiar sawtooth pattern around the nominal coverage level. Typically the coverage of the 95% confidence (or posterior) interval is near nominal, with fluctuations near the boundary of the

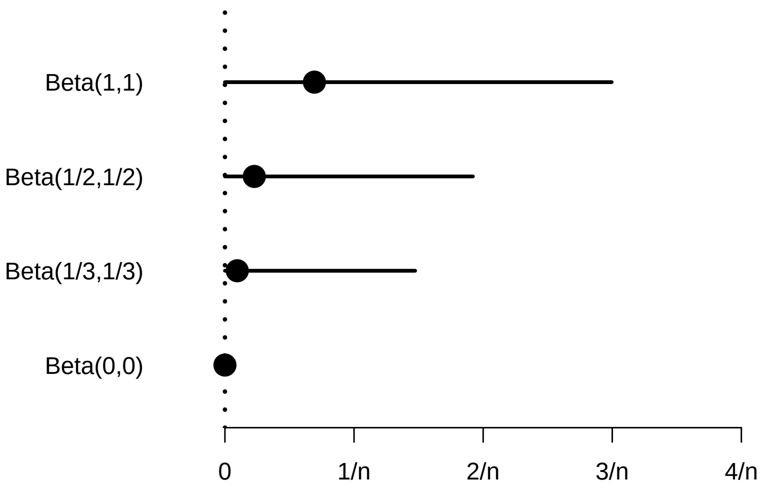

Fig 5. *One-sided posterior 95% intervals and medians (dots) for four conjugate beta models assuming different prior distributions Beta*(a,a) *for the extreme case of no observed successes* $(y=0)$ *in the limit as* $n\to\infty$. *In this case, since the likelihood does not contribute any information about the possibility of the rate or proportion being positive, all evidence to the contrary is supplied by the prior. This is evident when one lets a tend to zero as the posterior quantiles collapse all toward zero. This graph equivalently illustrates the corresponding gamma posteriors under four different priors Gamma*$(a,0)$ *when the exposure* X *equals some given number* $n>0$.

parameter space (see e.g., Agresti and Coull [1], Agresti and Min [2], Tuyl, Gerlach and Mengersen [20].) In the case of the binomial model, coverage probability plots show typically that near the boundaries, the coverage probability falls occasionally to as low as 80%, which is indeed the case for the Uniform prior. For the Jeffreys and the neutral priors, coverage probability appears to drop to 85% for a certain range of small values of θ. This may be somewhat disturbing, if one considers the minimum conditional coverage to be a decisive criterion for the goodness of an interval estimation procedure. However, the minimum coverage over the whole parameter space (excluding endpoints) is actually zero: the conditional coverage of the two-sided equal-tailed interval drops to zero eventually near the boundary, which is usually not evident in the plots. Under the uniform beta prior, the coverage drops to zero approximately for $\theta\leq 0.025/n$ for sample sizes (n) larger than 20. For the Neutral beta prior, the corresponding range is approximately $\theta\leq 0.00001/n$. It is of course possible to argue that these rates are negligibly small, but this argument should be irrelevant to a frequentist who assumes that prior knowledge of the true rate should be irrelevant.

Most practitioners are however probably more interested in the *average* performance rather than the *worst possible* performance over a sequence of trials [1, 3]. For example, for a regulatory authority monitoring an actual sequence of clinical trials, average performance over a long period of time should arguably

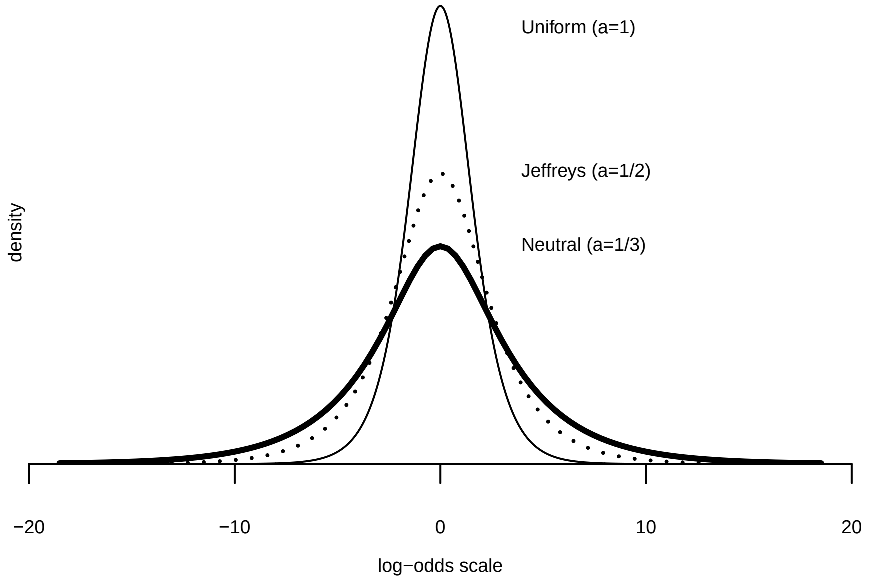

Fig 6. *Logistic (log-odds) transform $f(p) = \log(p/(1-p))$ of the three prior distributions Beta(a, a) with $a = 1$ (Uniform prior), $a = 1/2$ (Jeffreys prior) and $a = 1/3$ (Neutral prior). Compared to the uniform prior, Neutral and Jeffreys priors have heavier tails, explaining why there is less shrinkage of the posterior mass toward the prior median. The transformed Neutral prior is closely approximated between $(-20, 20)$ by a zero-centered t distribution with scale $10/3$ and $10/3$ degrees of freedom.*

be a more relevant criterion of the goodness of the interval estimation procedure. After all, the interest is to assess the state of nature, which is essentially a Bayesian concept; the relevant quantity is therefore a Bayesian average over the conditional coverage probabilities [3]. If the sequence of parameters emerge from the prior itself, that is, $\theta^{(i)} \sim \text{Beta}(a, b)$, say, it is then straightforward to see that a Bayesian $100(1-\alpha)\%$ posterior credible interval has an exact long-run frequentist coverage for all α and all sample sizes n [18].

Thus, we could for example choose a Beta(a, a) distribution to represent the prior distribution of θ and obtain nominal average coverage. From our perspective, assuming that binomial (clinical) trials come on average from the uniform distribution in the probability scale is unrealistic as there is a lot of prior weight on the proportions and rates near $\theta = 1/2$. It should be more plausible that the long-term sequence of trials may have much smaller rates. In this context, it may be enlightening to have a look at the prior density on a different scale. Figure 6 shows the Neutral, Jeffreys, and the Uniform priors translated into the logit (log-odds) scale. The smaller the shape parameter a is, the wider the density appears; the Neutral prior, with the most prior variance, also appears to have a heavy-tailed distribution resembling a t distribution. In contrast, the uniform prior appears to be relatively thin-tailed with only 1.3% of the mass outside the interval $(-5, 5)$, compared with 21.3% of the Neutral prior.

However in practice, small differences in coverage probabilities do not matter. At least in the case of clinical trials, where the long-run-frequentist performance is important, the sponsor always computes the performance of the pre-defined Bayesian decision rule under various assumed scenarios. If the decision rule yields too high a false positive rate, the rule is adjusted such that frequentist performance is acceptable. For example, in an early phase proof-of-concept trial, a drug may be considered promising if the posterior probability of the responder rate being greater than a given threshold τ is larger than a given "level of proof" probability L: $\Pr(\theta > \tau|y) \geq L$. If the operating characteristics of this rule are unsatisfactory, L can be raised to produce a lower false positive error [16]. Moreover, decision rules often involve two or more simultaneous rules so the false positive rate can often only be computed by simulation. It is therefore more important that the performance of the rule itself performs satisfactorily, under any given hypothetical outcome of a single trial. Focusing solely on the long-term averages is not desirable, as the actual decisions will be made, in the end, at the study level.

4.3. Point estimates

Under the $\text{Beta}(1/3, 1/3)$ prior, the posterior mean is $(y + 1/3)/(n + 2/3)$, with the weight on y/n larger than that of the Uniform and Jeffreys priors so that the posterior mean estimate is closer to the sample mean and less influenced by the implicit bias toward the value $1/2$. In practice, the common consensus tends to be that the true rate is expected to be close to the observed rate y/n. The Uniform distribution produces a posterior mean estimate of $(y+1)/(n+2)$, which may look suspiciously high, if n is large and y is small.

For comparison, consider also the Wilson point estimate $\hat{p} = (y + 2)/(n + 4)$ suggested by Agresti and Coull [1]. This point estimate improves the frequentist coverage probability of the Wald confidence interval that depends on $\hat{p}$. Although the starting point in the paper is not Bayesian, this estimate is equivalent to the posterior mean obtained under the prior $\text{Beta}(2, 2)$, which shrinks the point estimate strongly toward $1/2$, reflecting fairly high prior information (or, belief) about the true rate being around the middle of the parameter space. However in the analysis of rare events (especially, rare adverse events in clinical trials), a considerable departure from the observed point estimate may be controversial: the Wilson estimate may be almost three times as large as the observed one ($1/n$ vs. $3/(n + 4) \approx 3/n$).

4.4. Analysis of rare events

Due to its consistent behavior also for small values of y, the Neutral prior should be a suitable choice for the (default) analysis of rare events as well. It is however quite valid to ask, why would one ever use a symmetric prior for the analysis of rare events, as there would be an equal prior chance that the rate is extremely high? It is important to understand that a *default* prior exists to represent the uncertainty about *all* possible proportions and rates under the assumption of no specific prior knowledge of the magnitude. Once information is available (such as “the rate should be about 1/1000”), a symmetric default prior will obviously not be the best choice, as it turns out to be a skeptical (or, ‘conservative’) one as it assigns (some) weight to the possibility that the rate is actually very high. Still, the Neutral prior gives only little posterior weight to the possibility that the true rate is over 0.50, say: if one observes 0 events out of a sample size of three, the posterior probability that $\theta > 0.5$ is as small as 0.02.

It may be even be controversial to specify an informative prior when analyzing rare events, for example when analyzing drug safety data and adverse event counts. The Neutral prior provides a slightly skeptical default choice without excessive shrinkage; if an asymmetric prior is however preferred, one can use a weakly informative prior of the form (1) with a relatively small prior sample size m, perhaps even as small as $m = 1$. Using such a prior should not cause any doubts about lack of objectivity in the context of any properly powered trial.

5. Examples

5.1. *Weakly informative beta priors*

A single rare event was observed in a clinical trial with a sample size of 100, so that the observed rate was $p = 0.01$. A weakly informative prior is constructed, with a prior sample size $m = 1$ based on this information. The distributional form (1) with $a = 1/3$, $m = 1$, and $p = 0.01$, yields the prior $\mathrm{Beta}(1/3 + 0.01, 1/3 + 0.99)$. This is an extremely weakly informative prior as the implied number of prior successes is practically zero; consequently the default prior information overwhelms the information injected into the prior. A priori, $\Pr(\theta > p) = 0.77$, showing that the prior is skewed toward the prior median 0.5. Now, if we observe exactly the same result ($y = 1$ and $n = 100$) the posterior tail probability $\Pr(\theta > 0.01|y) = 0.508$, nearly equal to what one would expect from the Neutral prior in the case of observing one single success out of $n + 1$.

The prior, with $mp \approx 0$, and $m(1 - p) \approx 1$, is therefore practically worth zero successes out of a sample size of one. This illustrates the fact that it is not possible to inject very little information (by choosing a relatively small m) in a beta prior while expecting that it contains a lot of information about the location of θ with respect to m. The more precise information we wish to encode in the prior, the larger the implied number of prior observations $y = mp$ must be; if we choose a very small p, we correspondingly need to increase m. If mp is close to zero, the posterior distribution will be vague about the location of θ. The inferences relative to the scale will be also considerably affected by the underlying prior providing the default information in the model: most of the information comes from the underlying prior and not from the data.

For comparison, another prior was constructed by matching the observed rate p =0.01 to the prior mean, obtaining the prior $\mathrm{Beta}(0.01, 0.09)$. This gives the prior tail probability $\Pr(\theta > 0.01|y) = 0.04$, and the posterior tail probability

$\Pr(\theta > 0.01|y) = 0.370$, assuming $y = 1$ and $n = 100$. This shows again how the underlying prior, now $\mathrm{Beta}(0, 0)$, dominates prior and posterior inferences, this time pulling the posterior mass toward zero instead of toward $1/2$.

5.2. Informative beta priors

In a clinical study, the observed adverse event rate in the placebo group was observed to be 2 out of 1,400 ($\hat{\theta} = 0.00143$). To use some of this information in a future study, an informative distribution was proposed by matching the mean of a beta distribution and inflating the prior variance by fixing the 95% prior quantile to $5\hat{\theta}$, similar to the method used by Winkler, Smith and Fryback [21], obtaining $\mathrm{Beta}(0.042, 29.38)$.

This prior was evaluated by supposing that an identical result would have been obtained ($y = 2, n = 1400$). The posterior $\mathrm{Beta}(2.042, 1429.38)$ has the median at $0.84\hat{\theta}$, and $\Pr(\theta > \hat{\theta}|y) = 0.407$, showing that the prior still favors values near zero although the sample size of the trial was almost 50 times larger than that implied in the prior. This shows that the implicit prior distribution in this case, $\mathrm{Beta}(0, 0)$, affected the posterior inferences. This is not surprising, since the posterior is essentially equivalent to one derived under the Haldane prior from an outcome of $y = 2$ and $n = 1400 + 30$.

Surprisingly, it was found that the prior distribution $\mathrm{Beta}(0.042, 29.38)$ is *not even a unique solution* to the given constraints: $\mathrm{Beta}(0.224, 156.4)$ has also mean $\hat{\theta}$ and the 95% quantile at $5\hat{\theta}$, although its implied sample size is considerably larger.

Further, if the 95% quantile had been fixed to be 2 times the mean instead, again two distributions would have emerged to satisfy the two constraints, the one implying $m = 12.7$ and the other implying a prior sample size of as much as $m = 2488.4$, almost 200 times larger than that of the first candidate prior.

This shows clearly that merely looking at the mean and the quantiles of a prior may not be helpful in constructing a prior nor in assessing how much information there is actually is in a prior distribution.

5.3. Informative gamma priors

In the case of the gamma-Poisson model, the actual number of observed events must be known, not just the observed rate, to get the prior precision right. Gelman et al. [9] introduce the gamma conjugate model with an example of asthma mortality rates, which are given as “around 0.6 per 100,000”. This information is translated as a prior $\mathrm{Gamma}(3, 5)$ by matching the mean of the distribution to 0.6, and then adjust the parameters to “prior knowledge about the tail of the distribution.” The quoted observed rate does not explicitly tell us whether there were 6, 60, or 600 events observed (or more). The phrases ‘6 out of 1,000,000’ and ‘60 out of 10,000,000’ imply completely different posteriors if taken literally, since the larger the y, the larger the precision relative to the scale.

Assuming that there were $y = 6$ events in a million, an alternative prior Gamma$(1/3 + 6, 10)$ is set up. Suppose now, as in the book [9], we observe $y = 3$ over an exposure of 200,000 ($X = 2.0$), that is, an observed rate of 1.5 in 100,000. Under the original prior, we obtain the posterior Gamma$(3 + 3, 5 + 2)$ with median 0.81 and mean 0.86, a "considerable" shrinkage toward the prior. This should be not surprising, since the prior weight is $w = 5/(5 + 2) = 0.71$, so the effect of the data is relatively weak. Under the alternative prior, the posterior Gamma$(1/3 + 6 + 3, 10 + 2)$ has a median at 0.75, at the weighted average of the observed results. The prior weight is relatively large ($w = 10/(10 + 2) = 0.83$) also in this case.

A more robust prior could have been obtained by down-weighting the number of observations while keeping the median at approximately 0.6. Had a weakly informative prior Gamma$(1/3+0.6, 1)$ been used (that is, the information scaled by a factor of $k = 0.1$) yielding the posterior Gamma$(1/3 + 0.6 + 3, 1 + 2.0)$ the new information would have implied a posterior median of approximately 1.2, giving the prior much less weight, but still the relative weight on the prior median is quite large, w =1/(1+2)=1/3. Scaling the information by $k = 0.01$, we would have obtained a median of 1.46. This shows how sensitive the posterior can be to the choice of an informative prior, which should be chosen carefully so that the information obtained from a previous study does not have undue influence on the information on that obtained from the new study.

6. Conclusion

The implications of the choice of the prior should be made clear in advance so that all parties concerned are comfortable with the resulting posterior inferences before the data are analyzed. Often we simply look at the shape of the prior and possibly the implied prior sample size to assess the amount of information in the prior distribution, unaware of the behavior of the posterior when the prior is combined with the likelihood. It is however important to assess the effect of the prior by examining the implied posteriors under hypothetical outcomes. If under a supposedly noninformative or an informative prior, the posterior distribution under some hypothetical outcomes seem counterintuitive, this would be indicative of using a prior that conflicts with the (implicit) prior knowledge. This applies not only to conjugate priors but of course to other models as well.

Under the Neutral priors, the behavior of the models is consistent and intuitive in the sense that the center of the posterior mass is determined as the approximate weighted average of the prior median and the sample mean. Under the default noninformative prior, the prior median has weight zero, and the sample mean (the maximum likelihood estimate) has consequently an intuitive Bayesian interpretation as the approximate posterior median. I suggest these priors as default priors for applied Bayesian analyses.

Still, why should one expect a 50–50 split around the maximum likelihood estimate? One can as well ask: why should we accept a 25–75 split instead? It must be emphasized that it is not possible to claim that having the MLE at the median is “ideal”; it is rather a compromise that nevertheless is intuitively appealing for a default prior.

The neutral priors can be applied to other one-parameter conjugate models as well, provided that the likelihood implies a Beta$(a+r, a+s)$ posterior distribution (with the maximum likelihood estimate $r/(r+s)$), or a Gamma$(c+r, s)$ posterior distribution (with the maximum likelihood estimate r/s). For example, they can be used in conjunction with the negative binomial model and the exponential time-to-event model.

The method introduced for constructing informative priors should be useful whether or not one wishes to adopt a neutral noninformative prior as a default prior: regardless of the default prior adopted, it is suggested to encode information for these models using the distributional forms (1) or (2), in terms of the quantities (p, m) or (y, X, k). These distributional forms should help to assess the shrinkage of the median, and should also be useful when evaluating the effect of any beta and gamma distribution. No matter which prior distribution is used, one must be aware of the effect of the underlying default prior and be consistent when constructing informative priors by always keeping this one prior in the background in all informative priors that are constructed. A default prior is, essentially, ‘the prior of all priors.’ The shape parameters of a default prior could therefore be considered as an inherent part of the prior distribution, and not counted as part of the implied prior successes and failures; hence one could define the Neutral beta distribution as NeutralBeta$(a, b) \equiv$ Beta$(1/3+a, 1/3+b)$, where $a+b$ would be zero in the noninformative case and positive otherwise. Similarly, one could define a Neutral gamma distribution.

Acknowledgements

I thank Andrew Gelman, Ekkehard Glimm, Byron Jones, Peter Müller, Amy Racine, and the three anonymous referees for comments and suggestions.

References

[1] Agresti, A. and Coull, B. A. (1998). Approximate Is Better than ”Exact” for Interval Estimation of Binomial Proportions. *The American Statistician* **52** 119–126. MR1628435

[2] Agresti, A. and Min, Y. (2001). On Small-Sample Confidence Intervals for Parameters in Discrete Distributions. *Biometrics* **57** 963–971. MR1863460

[3] Bayarri, M. J. and Berger, J. O. (2004). The Interplay of Bayesian and Frequentist Analysis. *Statistical Science* **19** pp. 58-80. MR2082147

[4] Berg, C. and Pedersen, H. L. (2006). The Chen-Rubin Conjecture in a Continuous Setting. *Methods and Applications of Analysis* **13** 63–88. MR2275872

[5] Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* **1** 385–402. MR2221271

[6] Bernardo, J. M. (1979). Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)* **41** 113–147. MR0547240

[7] Bernardo, J. M. (2005). Reference Analysis. In *Handbook of Statistics 25: Bayesian Thinking, Modeling and Computation (D. K. Dey and C. R. Rao, eds.)* 17–90. Elsevier, Amsterdam. MR2490522

[8] Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Data Analysis*, 1st ed. Wiley-Interscience, New York.

[9] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, London. MR2027492

[10] Gelman, A., Jakulin, A., Pittau, M. G. and Yu, S.-S. (2008). A weakly informative default prior distrbution for logistic and other regression models. *The Annals of Applied Statistics* **2** 1360–1383. MR2655663

[11] Haldane, J. B. S. (1948). The Precision of Observed Values of Small Frequencies. *Biometrika* **35** 297–300. MR0029137

[12] Hanley, J. A. and Lippman-Hand, A. (1983). If nothing goes wrong, is everything all right? Interpreting zero numerators. *Journal of the American Medical Association* **249** 1743–1745.

[13] Jeffreys, H. (1961). *Theory of probability*, 3rd ed. Oxford University Press, New York. MR0187257

[14] Jovanovic, B. D. and Levy, P. S. (1997). A Look at the Rule of Three. *The American Statistician* **51** 137–139.

[15] Kerman, J. (2011). A closed-form approximation for the median of the beta distribution. arXiv:1111.0433v1 [math.ST]

[16] Neuenschwander, B., Rouyrre, N., Hollaender, N., Zuber, E. and Branson, M. (2011). A proof of concept phase II non-inferiority criterion. *Statistics in Medicine* **30** 1618–1627.

[17] O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and Rakow, T. (2006). *Uncertain judgements: Eliciting experts' Probabilities*. Wiley, Hoboken, NJ.

[18] Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics* **12** pp. 1151-1172. MR0760681

[19] Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, Chichester.

[20] Tuyl, F., Gerlach, R. and Mengersen, K. (2008). A comparison of Bayes-Laplace, Jeffreys, and other priors: the case of zero events. *The American Statistician* **62** 40–44. MR2416895

[21] Winkler, R. L., Smith, J. E. and Fryback, D. G. (2002). The Role of Informative Priors in Zero-Numerator Problems: Being Conservative versus Being Candid. *The American Statistician* **56** 1–4. MR1939390