5

Prior distributions

The prior distribution plays a defining role in Bayesian analysis. In view of the controversy surrounding its use it may be tempting to treat it almost as an embarrassment and to emphasise its lack of importance in particular applications, but we feel it is a vital ingredient and needs to be squarely addressed. In this chapter we introduce basic ideas by focusing on single parameters, and in subsequent chapters consider multi-parameter situations and hierarchical models. Our emphasis is on understanding what is being used and being aware of its (possibly unintentional) influence.

5.1 Different purposes of priors

A basic division can be made between so-called "non-informative" (also known as "reference" or "objective") and "informative" priors. The former are intended for use in situations where scientific objectivity is at a premium, for example, when presenting results to a regulator or in a scientific journal, and essentially means the Bayesian apparatus is being used as a convenient way of dealing with complex multi-dimensional models. The term "non-informative" is misleading, since all priors contain some information, so such priors are generally better referred to as "vague" or "diffuse." In contrast, the use of informative prior distributions explicitly acknowledges that the analysis is based on more than the immediate data in hand whose relevance to the parameters of interest is modelled through the likelihood, and also includes a considered judgement concerning plausible values of the parameters based on external information.

In fact the division between these two options is not so clear-cut — in particular, we would claim that any "objective" Bayesian analysis is a lot more "subjective" than it may wish to appear. First, any statistical model (Bayesian or otherwise) requires qualitative judgement in selecting its structure and distributional assumptions, regardless of whether informative prior distributions are adopted. Second, except in rather simple situations there may not be an agreed "objective" prior, and apparently innocuous assumptions can strongly influence conclusions in some circumstances.

In fact a combined strategy is often reasonable, distinguishing parameters of

primary interest from those which specify secondary structure for the model. The former will generally be location parameters, such as regression coefficients, and in many cases a vague prior that is locally uniform over the region supported by the likelihood will be reasonable. Secondary aspects of a model include, say, the variability between random effects in a hierarchical model. Often there is limited evidence in the immediate data concerning such parameters and hence there can be considerable sensitivity to the prior distribution, in which case we recommend thinking carefully about reasonable values in advance and so specifying fairly informative priors — the inclusion of such external information is unlikely to bias the main estimates arising from a study, although it may have some influence on the precision of the estimates and this needs to be carefully explored through sensitivity analysis. It is preferable to construct a prior distribution on a scale on which one has has a good interpretation of magnitude, such as standard deviation, rather than one which may be convenient for mathematical purposes but is fairly incomprehensible, such as the logarithm of the precision. The crucial aspect is not necessarily to avoid an influential prior, but to be aware of the extent of the influence.

5.2 Vague, "objective," and "reference" priors

5.2.1 Introduction

The appropriate specification of priors that contain minimal information is an old problem in Bayesian statistics: the terms "objective" and "reference" are more recent and reflect the aim of producing a baseline analysis from which one might possibly measure the impact of adopting more informative priors. Here we illustrate how to implement standard suggestions with BUGS. Using the structure of graphical models, the issue becomes one of specifying appropriate distributions on "founder" nodes (those with no parents) in the graph.

We shall see that some of the classic proposals lead to "improper" priors that do not form distributions that integrate to 1: for example, a uniform distribution over the whole real line, no matter how small the ordinate, will still have an infinite integral. In many circumstances this is not a problem, as an improper prior can still lead to a proper posterior distribution. BUGS in general requires that a full probability model is defined and hence forces all prior distributions to be proper — the only exception to this is the `dflat()` distribution (Appendix C.1). However, many of the prior distributions used are "only just proper" and so caution is still required to ensure the prior is not having unintended influence.

5.2.2 Discrete uniform distributions

For discrete parameters it is natural to adopt a discrete uniform prior distribution as a reference assumption. We have already seen this applied to the degrees of freedom of a t-distribution in Example 4.1.2, and in §9.8 we will see how it can be used to perform a non-Bayesian bootstrap analysis within BUGS.

5.2.3 Continuous uniform distributions and Jeffreys prior

When it comes to continuous parameters, it is tempting to automatically adopt a uniform distribution on a suitable range. However, caution is required since a uniform distribution for θ does not generally imply a uniform distribution for functions of θ. For example, suppose a coin is known to be biased, but you claim to have "no idea" about the chance θ of it coming down heads and so you give θ a uniform distribution between 0 and 1. But what about the chance (θ^2) of it coming down heads in both of the next two throws? You have "no idea" about that either, but according to your initial uniform distribution on θ, $\psi = \theta^2$ has a density $p(\psi) = 1/(2\sqrt{\psi})$, which can be recognised to be a Beta(0.5, 1) distribution and is certainly not uniform.

Harold Jeffreys came up with a proposal for prior distributions which would be invariant to such transformations, in the sense that a "Jeffreys" prior for θ would be formally compatible with a Jeffreys prior for any 1–1 transformation $\psi = f(\theta)$. He proposed defining a "minimally informative" prior for θ as $p_J(\theta) \propto I(\theta)^{1/2}$ where $I(\theta) = -E[\frac{d^2}{d\theta^2} \log p(Y|\theta)]$ is the Fisher information for θ (§3.6.1). Since we can also express $I(\theta)$ as

$$I(\theta) = E_{Y|\theta}\left[\left(\frac{d \log p(Y|\theta)}{d\theta}\right)^2\right],$$

we have

$$I(\psi) = I(\theta)\left|\frac{d\theta}{d\psi}\right|^2.$$

Jeffreys' prior is therefore invariant to reparameterisation since

$$I(\psi)^{1/2} = I(\theta)^{1/2}\left|\frac{d\theta}{d\psi}\right|,$$

and the Jacobian terms cancel when transforming variables via the expression in §2.4. Hence, a Jeffreys prior for θ transforms to a Jeffreys prior for any 1–1 function $\psi(\theta)$.

As an informal justification, Fisher information measures the curvature of the log-likelihood, and high curvature occurs wherever small changes in parameter values are associated with large changes in the likelihood: Jeffreys' prior gives more weight to these parameter values and so ensures that the

influence of the data and the prior essentially coincide. We shall see examples of Jeffreys priors in future sections.

Finally, we emphasise that if the specific form of vague prior is influential in the analysis, this strongly suggests you have insufficient data to draw a robust conclusion based on the data alone and that you should not be trying to be "non-informative" in the first place.

5.2.4 Location parameters

A location parameter θ is defined as a parameter for which $p(y|\theta)$ is a function of $y - \theta$, and so the distribution of $y - \theta$ is independent of θ. In this case Fisher's information is constant, and so the Jeffreys procedure leads to a uniform prior which will extend over the whole real line and hence be improper. In BUGS we could use `dflat()` to represent this distribution, but tend to use proper distributions with a large variance, such as `dunif(-100,100)` or `dnorm(0,0.0001)`: we recommend the former with appropriately chosen limits, since explicit introduction of these limits reminds us to be wary of their potential influence. We shall see many examples of this use, for example, for regression coefficients, and it is always useful to check that the posterior distribution is well away from the prior limits.

5.2.5 Proportions

The appropriate prior distribution for the parameter θ of a Bernoulli or binomial distribution is one of the oldest problems in statistics, and here we illustrate a number of options. First, both Bayes (1763) and Laplace (1774) suggest using a uniform prior, which is equivalent to Beta$(1, 1)$. A major attraction of this assumption, also known as the Principle of Insufficient Reason, is that it leads to a discrete uniform distribution for the predicted number y of successes in n future trials, so that $p(y) = 1/(n+1)$, $y = 0, 1, ..., n$,* which seems rather a reasonable consequence of "not knowing" the chance of success. On the $\phi = \text{logit}(\theta)$ scale, this corresponds to a standard logistic distribution, represented as `dlogis(0,1)` in BUGS (see code below).

Second, an (improper) uniform prior on ϕ is formally equivalent to the (improper) Beta$(0, 0)$ distribution on the θ scale, i.e., $p(\theta) \propto \theta^{-1}(1-\theta)^{-1}$: the code below illustrates the effect of bounding the range for ϕ and hence making these distributions proper. Third, the Jeffreys principle leads to a Beta$(0.5, 0.5)$ distribution, so that $p_J(\theta) = \pi^{-1}\theta^{\frac{1}{2}}(1-\theta)^{\frac{1}{2}}$. Since it is common to use normal prior distributions when working on a logit scale, it is of interest to consider what normal distributions on ϕ lead to a "near-uniform"

*See Table 3.1 — the posterior predictive distribution for a binomial observation and beta prior is a beta-binomial distribution. With no observed data, $n = y = 0$ in Table 3.1, this posterior predictive distribution becomes the *prior predictive* distribution, which reduces to the discrete uniform for $a = b = 1$.

distribution on θ. Here we consider two possibilities: assuming a prior variance of 2 for ϕ can be shown to give a density for θ that is "flat" at $\theta = 0.5$, while a normal with variance 2.71 gives a close approximation to a standard logistic distribution, as we saw in Example 4.1.1.

```
theta[1]          ~ dunif(0,1)     # uniform on theta
phi[1]            ~ dlogis(0,1)

phi[2]            ~ dunif(-5,5)    # uniform on logit(theta)
logit(theta[2]) <- phi[2]

theta[3]          ~ dbeta(0.5,0.5) # Jeffreys on theta
phi[3]           <- logit(theta[3])

phi[4]            ~ dnorm(0,0.5)   # var=2, flat at theta = 0.5
logit(theta[4]) <- phi[4]

phi[5]            ~ dnorm(0,0.368) # var=2.71, approx. logistic
logit(theta[5]) <- phi[5]
```

We see from Figure 5.1 that the first three options produce apparently very different distributions for θ, although in fact they differ at most by a single implicit success and failure (§5.3.1). The normal prior on the logit scale with variance 2 seems to penalise extreme values of θ, while that with variance 2.71 seems somewhat more reasonable. We conclude that, in situations with very limited information, priors on the logit scale could reasonably be restricted to have variance of around 2.7.

Example 5.2.1. *Surgery (continued): prior sensitivity*
What is the sensitivity to the above prior distributions for the mortality rate in our "Surgery" example (Example 3.3.2)? Suppose in one case we observe 0/10 deaths (Figure 5.2, left panel) and in another, 10/100 deaths (Figure 5.2, right panel). For 0/10 deaths, priors 2 and 3 pull the estimate towards 0, but the sensitivity is much reduced with the greater number of observations.

5.2.6 Counts and rates

For a Poisson distribution with mean θ, the Fisher information is $I(\theta) = 1/\theta$ and so the Jeffreys prior is the improper $p_J(\theta) \propto \theta^{-\frac{1}{2}}$, which can be approximated in BUGS by a `dgamma(0.5, 0.00001)` distribution. The same prior is appropriate if θ is a rate parameter per unit time, so that $Y \sim \text{Poisson}(\theta t)$.

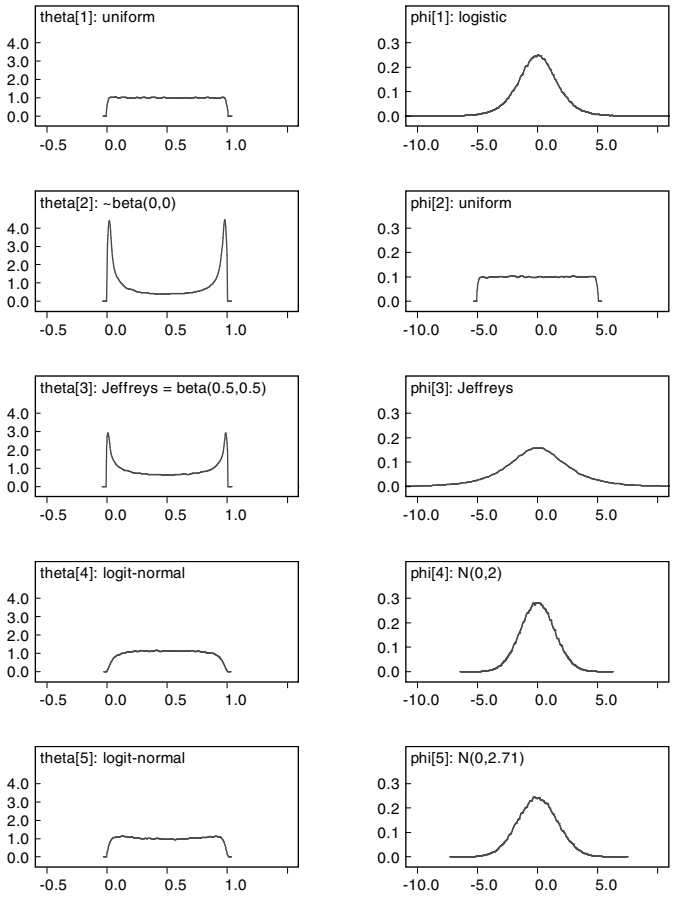

FIGURE 5.1
Empirical distributions (based on 100,000 samples) corresponding to various different priors for a proportion parameter.

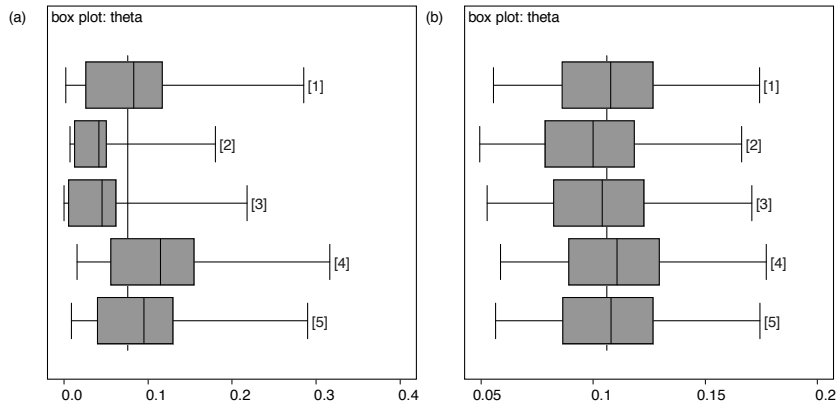

FIGURE 5.2
Box plots comparing posterior distributions arising from the five priors discussed above for mortality rate: (a) 0/10 deaths observed; (b) 10/100 deaths observed.

5.2.7 Scale parameters

Suppose σ is a scale parameter, in the sense that $p(y|\sigma) = \sigma^{-1} f(y/\sigma)$ for some function f, so that the distribution of Y/σ does not depend on σ. Then it can be shown that the Jeffreys prior is $p_J(\sigma) \propto \sigma^{-1}$, which in turn means that $p_J(\sigma^k) \propto \sigma^{-k}$, for any choice of power k. Thus for the normal distribution, parameterised in BUGS in terms of the precision $\tau = 1/\sigma^2$, we would have $p_J(\tau) \propto \tau^{-1}$. This prior could be approximated in BUGS by, say, a `dgamma(0.001,0.001)`, which also can be considered an "inverse-gamma distribution" on the variance σ^2. Alternatively, we note that the Jeffreys prior is equivalent to $p_J(\log \sigma^k) \propto$ const, i.e., an improper uniform prior. Hence it may be preferable to give $\log \sigma^k$ a uniform prior on a suitable range, for example, `log.tau ~ dunif(-10, 10)` for the logarithm of a normal precision. We would usually want the bounds for the uniform distribution to have negligible influence on the conclusions.

We note some potential conflict in our advice on priors for scale parameters: a uniform prior on $\log \sigma$ follows Jeffreys' rule but a uniform on σ is placing a prior on an interpretable scale. There usually would be negligible difference between the two — if there is a noticeable difference, then there is clearly little information in the likelihood about σ and we would recommend a weakly informative prior on the σ scale.

Note that the advice here applies only to scale parameters governing the variance or precision of *observable* quantities. The choice of prior for the variance of *random effects* in a hierarchical model is more problematic — we discuss this in §10.2.3.

5.2.8 Distributions on the positive integers

Jeffreys (1939) [p. 238] suggested that a suitable prior for a parameter N, where $N = 0, 1, 2, ...$, is $p(N) \propto 1/N$, analogously to a scale parameter.

Example 5.2.2. *Coin tossing: estimating number of tosses*
Suppose we are told that a fair coin has come up heads $y = 10$ times. How many times has the coin been tossed? Denoting this unknown quantity by N we can write down the likelihood as

$$p(y|N) = \text{Binomial}(0.5, N) \propto \frac{N!}{(N-y)!}0.5^N.$$

As N is integer-valued we must specify a *discrete* prior distribution.

Suppose we take Jeffreys' suggestion and assign a prior $p(N) \propto 1/N$, which is improper but could be curtailed at a very high value. Then the posterior distribution is

$$p(N|y) \propto \frac{N!}{(N-y)!}0.5^N/N \propto \frac{(N-1)!}{(N-y)!}0.5^N, \;\; N \geq y,$$

which we can recognise as the kernel of a negative binomial distribution with mean $2y = 20$. This has an intuitive attraction, since if instead we had fixed $y = 10$ in advance and flipped a coin until we had y heads, then the sampling distribution for the random quantity N would be just this negative binomial. However, it is notable that we were *not* told that this was the design — we have no idea whether the final flip was a head or not.

Alternatively, we may wish to assign a uniform prior over integer values from 1 to 100, i.e., $\Pr(N = n) = 1/100$, $n = 1, ..., 100$. Then the posterior for N is proportional to the likelihood, and its expectation, for example, is given by

$$E[N|y] = \sum_{n=1}^{100} n \Pr(N = n|y) = A \sum_{n=1}^{100} \frac{n \times n!}{(n-y)!}0.5^n, \tag{5.1}$$

where A is the posterior normalising constant. The right-hand side of (5.1) cannot be simplified analytically and so is cumbersome to evaluate (although this is quite straightforward with a little programming). In BUGS we simply specify the likelihood and the prior as shown below.

```
y <- 10
y  ~ dbin(0.5, N)
N  ~ dcat(p[])
for (i in 1:100) {p[i] <- 1/100}
```

BUGS can use the resulting samples to summarise the posterior graphically as well as numerically. Numeric summaries, such as the one shown below, allow us to make formal inferences; for example, we can be 95% certain that the coin has been tossed between 13 and 32 times. Graphical summaries, on the other hand,

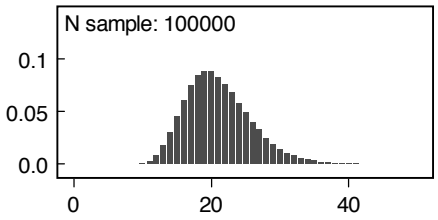

FIGURE 5.3
Approximate posterior distribution for number of (unbiased) coin tosses leading to ten heads.

might reveal interesting features of the posterior. Figure 5.3 shows the posterior density for N. Note that the mode is 20, which is the intuitive answer, as well as being the MLE and the posterior mean using the Jeffreys prior. Note also that although the uniform prior supports values in $\{1, ..., 9\}$, which are impossible in light of the observed data (10 heads), the posterior probability for these values is, appropriately, zero.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
N	21.01	4.702	0.01445	13.0	20.0	32.0	1	100000

In Example 5.5.2 we consider a further example of a prior over the positive integers which reveals the care that can be required.

5.2.9 More complex situations

Jeffreys' principle does not extend easily to multi-parameter situations, and additional context-specific considerations generally need to be applied, such as assuming prior independence between location and scale parameters and using the Jeffreys prior for each, or specifying an ordering of parameters into groups of decreasing interest.

5.3 Representation of informative priors

Informative prior distributions can be based on pure judgement, a mixture of data and judgement, or data alone. Of course, even the selection of relevant data involves a substantial degree of judgement, and so the specification of an informative prior distribution is never an automatic procedure.

We summarise some basic techniques below, emphasising the mapping of relevant data and judgement onto appropriate parametric forms, ideally representing "implicit" data.

5.3.1 Elicitation of pure judgement

Elicitation of subjective probability distributions is not a straightforward task due to a number of potential biases that have been identified. O'Hagan et al. (2006) provide some "*Guidance for best practice*," emphasising that probability assessments are constructed by the questioning technique, rather than being "pre-formed quantifications of pre-analysed belief" (p. 217). They say it is best to interview subjects face-to-face, with feedback and continual checking for biases, conducting sensitivity analysis to the consequence of the analysis, and avoiding verbal descriptions of uncertainty. They recommend eliciting intervals with moderate rather than high probability content, say by focusing on 33% and 67% quantiles: indeed one can simply ask for an interval and afterwards elicit a 'confidence' in that assessment (Kynn, 2005). They suggest using multiple experts and reporting a simple average, but it is also important to acknowledge imperfections in the process, and that even genuine "expertise" cannot guarantee a suitable subject. See also Kadane and Wolfson (1998) for elicitation techniques for specific models.

In principle any parametric distribution can be elicited and used in BUGS. However, it can be advantageous to use conjugate forms since, as we have seen in Chapter 3, the prior distribution can then be interpreted as representing "implicit data," in the sense of a prior estimate of the parameter and an "effective prior sample size." It might even then be possible to include the prior information as "data" and use standard classical methods (and software) for statistical analysis.

Below we provide a brief summary of situations: in each case the "implicit data" might be directly elicited, or measures of central tendency and spread requested and an appropriate distribution fitted. A simple moment-based method is to ask directly for the mean and standard deviation, or elicit an approximate 67% interval (i.e., the parameter is assessed to be twice as likely to be inside the interval as outside it) and then treat the interval as representing the mean $\pm$ 1 standard deviation, and solve for the parameters of the prior distribution. In any case it is good practice to iterate between alternative representations of the prior distribution, say as a drawn distribution, percentiles, moments, and interpretation as "implicit data," in order to check the subject is happy with the implications of their assessments.

- Binomial proportion θ. Suppose our prior information is equivalent to having observed y events in a sample size of n, and we wanted to derive a corresponding Beta(a, b) prior for θ. Combining an improper Beta(0,0) "pre-prior" with these implicit data gives a conjugate "posterior" of Beta$(y, n - y)$, which we can interpret as our elicited prior. The mean

of this elicited prior is $a/(a+b) = y/n$, the intuitive point estimate for θ, and the implicit sample size is $a+b = n$. Using a uniform "pre-prior" instead of the Beta(0,0) gives $a = y+1$ and $b = n-y+1$.

Alternatively, a moment-based method might proceed by eliciting a prior standard deviation as opposed to a prior sample size, and by then solving the mean and variance formulae (Appendix C.3) for a and b: $a = mb/(1-m)$, $b = m(1-m)^2/v + m - 1$, for an elicited mean $m = \hat{\theta}$ and variance v.

- Poisson rate θ: if we assume θ has a Gamma(a, b) distribution we can again elicit a prior estimate $\hat{\theta} = a/b$ and an effective sample size of b, assuming a Gamma(0,0) pre-prior (see Table 3.1, Poisson-gamma conjugacy), or we can use a moment-based method instead.

- Normal mean μ: a normal distribution can be obtained be eliciting a mean γ and standard deviation ω directly or via an interval. By conditioning on a sampling variance σ^2, we can calculate an effective prior sample size $n_0 = \sigma^2/\omega^2$ which can be fed back to the subject.

- Normal variance σ^2: $\tau = \sigma^{-2}$ may be assumed to have a Gamma(a, b) distribution, where a/b is set to an estimate of the precision, and $2a$ is the effective number of prior observations, assuming a Gamma(0,0) pre-prior (see Table 3.1, normal y with unknown variance σ^2).

- Regression coefficients: In many circumstances regression coefficients will be unconstrained parameters in standard generalised linear models, say log-odds ratios in logistic regression, log-rate-ratios in Poisson regression, log-hazard ratios in Cox regression, or ordinary coefficients in standard linear models. In each case it is generally appropriate to assume a normal distribution. Kynn (2005) described the elicitation of regression coefficients in GLMs by asking an expert for expected responses for different values of a predictor. Lower and upper estimates, with an associated degree of confidence, were also elicited, and the answers used to derive piecewise-linear priors.

Example 5.3.1. *Power calculations*
A randomised trial is planned with n patients in each of two arms. The response within each treatment arm is assumed to have between-patient standard deviation σ, and the estimated treatment effect Y is assumed to have a Normal$(\theta, 2\sigma^2/n)$ distribution. A trial designed to have two-sided Type I error α and Type II error β in detecting a true difference of θ in mean response between the groups will require a sample size per group of

$$n = \frac{2\sigma^2}{\theta^2}(z_{1-\beta} + z_{1-\alpha/2})^2,$$

where $\Pr(Z < z_p) = p$ for a standard normal variable $Z \sim \text{Normal}(0, 1)$. Alternatively, for fixed n, the power of the study is

$$\text{Power} = \Phi\left(\sqrt{\frac{n\theta^2}{2\sigma^2}} - z_{1-\alpha/2}\right).$$

If we assume $\theta = 5, \sigma = 10$, $\alpha = 0.05, \beta = 0.10$, so that the power of the trial is 90%, then we obtain $z_{1-\beta} = 1.28$, $z_{1-\alpha/2} = 1.96$, and $n = 84$.

Suppose we wish to acknowledge uncertainty about the alternative hypothesis θ and the standard deviation σ. First, we assume past evidence suggests θ is likely to lie anywhere between 3 and 7, which we choose to interpret as a 67% interval ($\pm$ 1 standard deviation), and so $\theta \sim \text{Normal}(5, 2^2)$. Second, we assess our estimate of $\sigma = 10$ as being based on around 40 observations, from which we assume a Gamma(a, b) prior distribution for $\tau = 1/\sigma^2$ with mean $a/b = 1/10^2$ and effective sample size $2a = 40$, from which we derive $\tau \sim \text{Gamma}(20, 2000)$.

```
tau    ~ dgamma(20, 2000)
sigma <- 1/sqrt(tau)
theta  ~ dnorm(5, 0.25)
n     <- 2*pow((1.28 + 1.96)*sigma/theta, 2)  # n for 90% power
power <- phi(sqrt(84/2)*theta/sigma - 1.96)   # power for n = 84
p70   <- step(power - 0.7)                    # Pr(power > 70%)
```

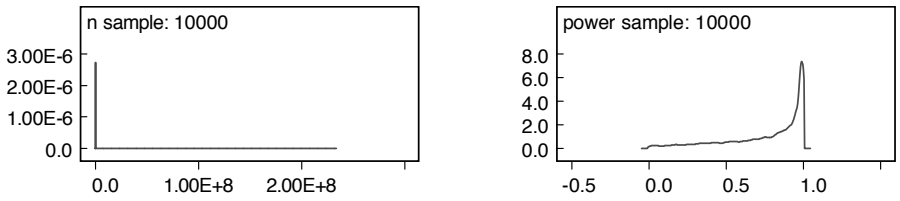

FIGURE 5.4
Empirical distributions based on 10,000 simulations for: `n`, the number of subjects required in each group to achieve 90% power, and `power`, the power achieved with 84 subjects in each group.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
n	38740.0	2.533E+6	25170.0	24.73	87.93	1487.0	1	10000
p70	0.7012	0.4577	0.004538	0.0	1.0	1.0	1	10000
power	0.7739	0.2605	0.002506	0.1151	0.8863	1.0	1	10000

Note that the median values for `n` (88) and `power` (0.89) are close to the values derived by assuming fixed θ and σ (84 and 0.90, respectively), but also note the

huge uncertainty. It is quite plausible, under the considered prior for θ and σ, that to achieve 90% power the trial may need to include nearly 3000 subjects. Then again, we might get away with as few as 50! A trial involving 84 subjects in each group could be seriously underpowered, with 12% power being quite plausible. Indeed, there is a 30% chance that the power will be less than 70%.

5.3.2 Discounting previous data

Suppose we have available some historical data and we could obtain a prior distribution for the parameter θ based on an empirical estimate $\hat{\theta}_H$, say, by matching the prior mean and standard deviation to $\hat{\theta}_H$ and its estimated standard error. If we were to use this prior directly then we would essentially be pooling the data in a form of meta-analysis (see §11.4), in which case it would be preferable (and essentially equivalent) to use a reference prior and include the historical data directly in the model.

If we are reluctant to do this, it must be because we do not want to give the historical data full weight, perhaps because we do not consider it to have the same relevance and rigour as the data directly being analysed. We may therefore wish to *discount* the historical data using one of the methods outlined below.

- *Power prior:* this uses a prior mean based on the historical estimate $\hat{\theta}_H$, but discounts the "effective prior sample size" by a factor κ between 0 and 1: for example, a fitted Beta(a, b) would become a Beta$(\kappa a, \kappa b)$, a Gamma(a, b) would become a Gamma$(\kappa a, \kappa b)$, a Normal(γ, ω^2) would become a Normal$(\gamma, \omega^2/\kappa)$ (Ibrahim and Chen, 2000).

- *Bias modelling:* This explicitly considers that the historical data may be biased, in the sense that the estimate $\hat{\theta}_H$ is estimating a slightly different quantity from the θ of current interest. We assume that $\theta = \theta_H + \delta$, where δ is the bias whose distribution needs to be assessed. We further assume $\delta \sim [\mu_\delta, \sigma_\delta^2]$, where $[,]$ indicates a mean and variance but otherwise unspecified distribution. Then if we assume the historical data give rise to a prior distribution $\theta_H \sim [\gamma_H, \omega_H^2]$, we obtain a prior distribution for θ of
$$\theta \sim [\gamma_H + \mu_\delta, \omega_H^2 + \sigma_\delta^2].$$
Thus the prior mean is shifted and the prior variance is increased.

The power prior only deals with variability — the discount factor κ essentially represents the "weight" on a historical observation, which is an attractive concept to communicate but somewhat arbitrary to assess. In contrast, the bias modelling approach allows biases to be added, and the parameters can be defined in terms of the size of potential biases.

Example 5.3.2. *Power calculations (continued)*
We consider the power example (Example 5.3.1) but with both prior distributions discounted. We assume each historical observation informing the prior distribution for σ is only worth half a current observation, so that the prior for σ is only based on 10 rather than 20 observations. This discounts the parameters in the gamma distribution for τ by a factor of 2. For the treatment effect, we assume that the historical experiment could have been more favourable than the current one, so that the historical treatment effect had a bias with mean -1 and SD 2, and so would be expected to be between -5 and 3. Thus an appropriate prior distribution is $\theta \sim \text{Normal}(5-1, 2^2+2^2)$ or $\text{Normal}(4, 8)$ — this has been constrained to be > 0 using the `I(,)` construct (see Appendix A.2.2 and §9.6). This leads to the code:

```
# tau    ~ dgamma(20, 2000)
  tau    ~ dgamma(10, 1000)      # discounted by 2
# theta  ~ dnorm(5, 0.25)
  theta  ~ dnorm(4, 0.125)I(0,)  # 4 added to var and shifted
                                 # by -1, constrained to be >0
```

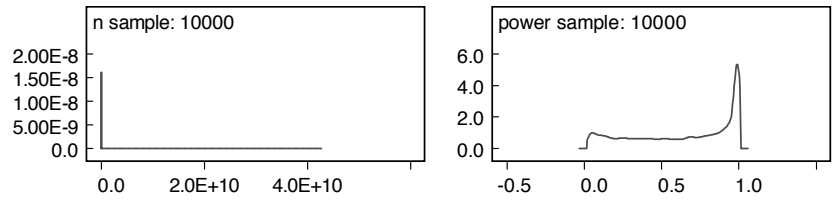

FIGURE 5.5
Empirical distributions based on 10,000 simulations for: `n`, the number of subjects required in each group to achieve 90% power, and `power`, the power achieved with 84 subjects in each group. Discounted priors for `tau` and `theta` used.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
n	4.542E+6	4.263E+8	4.26E+6	20.96	125.6	14270.0	1	10000
p70	0.5398	0.4984	0.005085	0.0	1.0	1.0	1	10000
power	0.6536	0.3315	0.003406	0.04353	0.7549	1.0	1	10000

This has raised the median sample size to 126, but with huge uncertainty. There is a 46% probability that the power is less than 70% if the sample size stays at 84.

5.4 Mixture of prior distributions

Suppose we want to express doubt about which of two or more prior distributions is appropriate for the data in hand. For example, we might suspect that *either* a drug will produce a similar effect to other related compounds, *or* if it doesn't behave like these compounds we are unsure about its likely effect.

For two possible prior distributions $p_1(\theta)$ and $p_2(\theta)$ for a parameter θ, the overall prior distribution is then a *mixture*

$$p(\theta) = qp_1(\theta) + (1-q)p_2(\theta),$$

where q is the assessed probability that p_1 is "correct." If we now observe data y, it turns out that the posterior for θ is

$$p(\theta|y) = q'p_1(\theta|y) + (1-q')p_2(\theta|y)$$

where

$$p_i(\theta|y) \propto p(y|\theta)p_i(\theta),$$
$$q' = \frac{qp_1(y)}{qp_1(y) + (1-q)p_2(y)},$$

where $p_i(y) = \int p(y|\theta)p_i(\theta)\,d\theta$ is the predictive probability of the data y assuming $p_i(\theta)$. The posterior is a mixture of the respective posterior distributions under each prior assumption, with the mixture weights adapted to support the prior that provides the best prediction for the observed data.

This structure is easy to implement in BUGS for any form of prior assumptions. We first illustrate its use with a simple example and then deal with some of the potential complexities of this formulation. In the example, `pick` is a variable taking the value j when the prior assumption j is selected in the simulation.

Example 5.4.1. *A biased coin?*
Suppose a coin is either unbiased or biased, in which case the chance of a "head" is unknown and is given a uniform prior distribution. We assess a prior probability of 0.9 that it is unbiased, and then observe 15 heads out of 20 tosses — what is the chance that the coin is biased?

```
r <- 15; n <- 20          # data
#####################################
r         ~ dbin(p, n)    # likelihood
p        <- theta[pick]
pick      ~ dcat(q[])     # 2 if biased, 1 otherwise
q[1]     <- 0.9
```

```
q[2]     <- 0.1
theta[1] <- 0.5           # if unbiased
theta[2]  ~ dunif(0, 1)   # if biased
biased   <- pick - 1      # 1 if biased, 0 otherwise
```

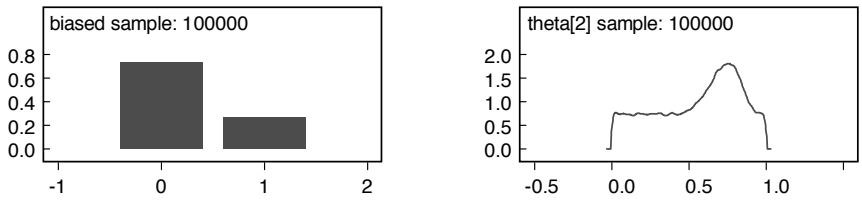

FIGURE 5.6
Biased coin: empirical distributions based on 100,000 simulations.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
biased	0.2619	0.4397	0.002027	0.0	0.0	1.0	1	100000
theta[2]	0.5594	0.272	9.727E-4	0.03284	0.6247	0.9664	1	100000

So the probability that the coin is biased has increased from 0.1 to 0.26 on the basis of the evidence provided. The rather strange shape of the posterior distribution for `theta[2]` is explained below.

If the alternative prior assumptions for `theta` in Example 5.4.1 were from the same parametric family, e.g., beta, then we could formulate this as `p` $\sim$ `dbeta(a[pick], b[pick])`, say, with specified values of `a[1]`, `a[2]`, `b[1]`, and `b[2]`. However, the more general formulation shown in the example allows prior assumptions of arbitrary structure.

It is important to note that when `pick=1`, `theta[1]` is sampled from its *posterior* distribution, but `theta[2]` is sampled from its *prior* as `pick=1` has essentially "cut" the connection between the data and `theta[2]`. At another MCMC iteration, we may have `pick=2` and so the opposite will occur, and this means that the posterior for each `theta[j]` recorded by BUGS is a mixture of "true" (model specific) posterior and its prior. This explains the shape of the posterior for `theta[2]` in the example above. If we are interested in the posterior distribution under each prior assumption individually, then we could do a separate run under each prior assumption, or only use those values for `theta[j]` simulated when `pick=j`: this "post-processing" would have to be performed outside BUGS.

We are essentially dealing with alternative model formulations, and our q's above correspond to posterior probabilities of models. There are well-known difficulties with these quantities both in theory, due to their potential

dependence on the within-model prior distributions, and in particular when calculating within MCMC: see §8.7. In principle we can use the structure above to handle a list of arbitrary alternative models, but in practice considerable care is needed if the sampler is not to go "off course" when sampling from the prior distribution at each iteration when that model is not being "picked." It is possible to define "pseudo-priors" for these circumstances, where `pick` also dictates the prior to be assumed for `theta[j]` when `pick` $\neq j$ — see §8.7 and Carlin and Chib (1995).

5.5 Sensitivity analysis

Given that there is no such thing as the *true* prior, sensitivity analysis to alternative prior assumptions is vital and should be an integral part of Bayesian analysis. The phrase "community of priors" (Spiegelhalter et al., 2004) has been used in the clinical trials literature to express the idea that different priors may reflect different perspectives: in particular, the concept of a "sceptical prior" has been shown to be valuable. Sceptical priors will typically be centred on a "null" value for the relevant parameter with the spread reflecting plausible but small effects. We illustrate the use of sceptical and other prior distributions in the following example, where the evidence for an efficacious intervention following myocardial infarction is considered under a range of priors for the treatment effect, namely, "vague," "sceptical," "enthusiastic," "clinical," and "just significant."

Example 5.5.1. *GREAT trial*

Pocock and Spiegelhalter (1992) examine the effect of anistreplase on recovery from myocardial infarction. 311 patients were randomised to receive either anistreplase or placebo (conventional treatment); the number of deaths in each group is given in the table below.

		Treatment		total
		anistreplase	placebo	
Event	death	13	23	36
	no death	150	125	275
total		163	148	311

Let r_j, n_j, and π_j denote the number of deaths, total number of patients, and underlying mortality rate, respectively, in group $j \in \{1, 2\}$ (1 = anistreplase; 2 = placebo). Inference is required on the log-odds ratio (log(OR)) for mortality in the anistreplase group compared to placebo, that is,

$$\delta = \log\left\{\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right\} = \text{logit}\,\pi_1 - \text{logit}\,\pi_2. \tag{5.2}$$

A classical maximum likelihood estimator and approximate variance are given by

$$\hat{\delta} = \log\left\{\frac{r_1/(n_1 - r_1)}{r_2/(n_2 - r_2)}\right\}, \quad V(\hat{\delta}) \approx s^2 = \frac{1}{r_1} + \frac{1}{r_2} + \frac{1}{n_1 - r_1} + \frac{1}{n_2 - r_2}.$$

For the above data these give $\hat{\delta} = -0.753$ with $s = 0.368$. An approximate Bayesian analysis might proceed via the assumption $\hat{\delta} \sim$ Normal(δ, s^2) with a locally uniform prior on δ, e.g., $\delta \sim$ Uniform$(-10, 10)$. A more appropriate likelihood is a binomial assumption for each observed number of deaths: $r_j \sim$ Binomial(π_j, n_j), $j = 1, 2$. In this case we could be "vague" by specifying Jeffreys priors for the mortality rates, $\pi_j \sim$ Beta$(0.5, 0.5)$, $j = 1, 2$, and then deriving the posterior for δ via (5.2). Alternatively we might parameterise the model directly in terms of δ:

$$\text{logit}\,\pi_1 = \alpha + \delta/2, \quad \text{logit}\,\pi_2 = \alpha - \delta/2,$$

which facilitates the specification of informative priors for δ. Here α is a nuisance parameter and is assigned a vague normal prior: $\alpha \sim$ Normal$(0, 100^2)$. Our first informative prior for δ is a "clinical" prior based on expert opinion: a senior cardiologist, informed by one unpublished and two published trials, expressed belief that *"an expectation of 15–20% reduction in mortality is highly plausible, while the extremes of no benefit and a 40% relative reduction are both unlikely."* This is translated into a normal prior with a 95% interval of -0.51 to 0 (0.6 to 1.0 on the OR scale): $\delta \sim$ Normal$(-0.26, 0.13^2)$. We also consider a "sceptical" prior, which is designed to represent a reasonable expression of doubt, perhaps to avoid early stopping of trials due to fortuitously positive results. For example, a hypothetical sceptic might find treatment effects more extreme than a 50% reduction or 100% increase in mortality largely implausible, giving a 95% prior interval (assuming normality) of -0.69 to 0.69 (0.5 to 2 on the OR scale): $\delta \sim$ Normal$(0, 0.35^2)$.

As a counterbalance to the sceptical prior we might specify an "enthusiastic" or "optimistic" prior, as a basis for conservatism in the face of early negative results, say. Such a prior could be centred around some appropriate beneficial treatment effect with a small prior probability (e.g., 5%) assigned to negative treatment benefits. We do not construct such a prior in this example, however, since the clinical prior described above also happens to be "enthusiastic" in this sense. Another prior of interest is the "just significant" prior. Assuming that the treatment effect is significant under a vague prior, it is instructive to ask how sceptical we would have to be for that significance to vanish. Hence we assume $\delta \sim$ Normal$(0, \sigma_\delta^2)$ and we search for the largest value of σ_δ such that the 95% posterior credible interval (just) includes zero. BUGS code for performing such a search is presented below along with code to implement the clinical, sceptical, and vague priors discussed above. (Note that a preliminary search had been run to identify the approximate value of σ_δ as somewhere between 0.8 and 1, though closed form approximations exist for this "just signficant" prior (Matthews, 2001; Spiegelhalter et al., 2004)).

```
model {
  for (i in 1:nsearch) {               # search for "just
    pr.sd[i]         <- start + i*step # significant" prior
    pr.mean[i]       <- 0
  }
  pr.mean[nsearch+1] <- -0.26
  pr.sd[nsearch+1]   <- 0.13           # clinical prior
  pr.mean[nsearch+2] <- 0
  pr.sd[nsearch+2]   <- 0.35           # sceptical prior

  # replicate data for each prior and specify likelihood...
  for (i in 1:(nsearch+3)) {
    for (j in 1:2) {
      r.rep[i,j]     <- r[j]
      n.rep[i,j]     <- n[j]
      r.rep[i,j]      ~ dbin(pi[i,j], n.rep[i,j])
    }
  }
  delta.mle          <- -0.753
  delta.mle           ~ dnorm(delta[nsearch+4], 7.40)

  # define priors and link to log-odds...
  for (i in 1:(nsearch+2)) {
    logit(pi[i,1])   <- alpha[i] + delta[i]/2
    logit(pi[i,2])   <- alpha[i] - delta[i]/2
    alpha[i]          ~ dnorm(0, 0.0001)
    delta[i]          ~ dnorm(pr.mean[i], pr.prec[i])
    pr.prec[i]       <- 1/pow(pr.sd[i], 2)
  }
  pi[nsearch+3,1]     ~ dbeta(0.5, 0.5)
  pi[nsearch+3,2]     ~ dbeta(0.5, 0.5) # Jeffreys prior
  delta[nsearch+3]   <- logit(pi[nsearch+3,1])
                      - logit(pi[nsearch+3,2])
  delta[nsearch+4]    ~ dunif(-10, 10) # locally uniform prior
}

list(r = c(13, 23), n = c(163, 148),
start = 0.8, step = 0.005, nsearch = 40)
```

The derived value of σ_δ is ~ 0.925, corresponding to the 25th element of `delta[]` above. Selected posterior and prior distributions are summarised below. We note the essentially identical conclusions of the classical maximum likelihood approach and the two analyses with vague priors. The results suggest we should conclude that anistreplase is a superior treatment to placebo if we are either (a priori) completely ignorant of possible treatment effect sizes, or we trust the senior cardiologist's expert opinion, or perhaps if we are otherwise enthusiastic about the

new treatment's efficacy. If, on the other hand, we wish to claim prior indifference as to the sign of the treatment effect but we believe "large" treatment effects to be implausible, we should be more cautious. The "just significant" prior has a 95% interval of $(\exp(-1.96 \times 0.925), \exp(1.96 \times 0.925)) = (0.16, 6.1)$ on the OR scale, corresponding to reductions/increases in mortality as extreme as 84%/610%. These seem quite extreme, implying that only a small degree of scepticism is required to render the analysis "non-significant." We might conclude that the GREAT trial alone does not provide "credible" evidence for superiority, and larger-scale trials are required to quantify the treatment effect precisely.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
delta[25]	-0.6635	0.3423	5.075E-4	-1.343	-0.6609	3.598E-4	1001	500000
delta[41]	-0.317	0.1223	1.741E-4	-0.5562	-0.317	-0.07745	1001	500000
delta[42]	-0.3664	0.2509	3.497E-4	-0.8608	-0.366	0.1245	1001	500000
delta[43]	-0.7523	0.367	5.342E-4	-1.487	-0.7479	-0.04719	1001	500000
delta[44]	-0.7534	0.3673	5.432E-4	-1.475	-0.7529	-0.0334	1001	500000

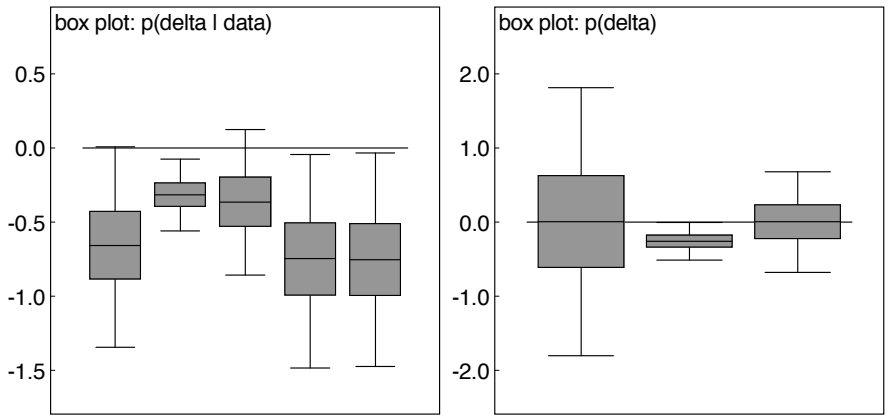

FIGURE 5.7
Left-hand side: Posterior distributions for δ from analysis of GREAT trial data. From left to right: corresponding to "just significant," "clinical," "sceptical," "Jeffreys" and "locally uniform" priors. Right-hand side: Prior distributions for analysis of GREAT trial data. From left to right: "just significant," "clinical" and "sceptical."

A primary purpose of trying a range of reasonable prior distributions is to find unintended sensitivity to apparently innocuous "non-informative" assumptions. This is reflected in the following example.

Example 5.5.2. *Trams: a classic problem from Jeffreys (1939)*
Suppose you enter a town of unknown size whose trams you know are numbered consecutively from 1 to N. You first see tram number $y = 100$. How large might N be?

We first note that the sampling distribution is uniform between 1 and N, so that $p(y|N) = \frac{1}{N}$, $y = 1, 2, \ldots, N$. Therefore the likelihood function for N is $\propto 1/N$, $N \geq y$, so that y maximises the likelihood function and so is the maximum likelihood estimator. The maximum likelihood estimate is therefore 100, which does not appear very reasonable.

Suppose we take a Bayesian approach and consider the prior distributions on the positive integers explored earlier (Example 5.2.2) — we will first examine the consequences using WinBUGS and then algebraically. We first consider a prior that is uniform on the integers up to an arbitrary upper bound M, say 5000. Y is assumed drawn from a categorical distribution: the following code shows how to set a uniform prior for N over the integers 1 to 5000 (as in Example 5.2.2) and how to use the `step` function to create a uniform sampling distribution between 1 and N.

```
Y <- 100
########################
Y            ~ dcat(p[])
# sampling distribution is uniform over first N integers
# use step function to change p[j] to 0 for j>N
for (j in 1:M) {
  p[j]      <- step(N - j + 0.01)/N
}
N            ~ dcat(p.unif[])
for (j in 1:M) {
  p.unif[j] <- 1/M
}
```

```
node mean   sd     MC error 2.5%  median 97.5%  start sample
N    1274.0 1295.0 10.86    109.0 722.0  4579.0 1001  10000
```

The posterior mean is 1274 and the median is 722, reflecting a highly skewed distribution. But is this a sensible conclusion? For an improper uniform prior over the whole of the integers, the posterior distribution is

$$p(N|y) \propto p(y|N)p(N) \propto 1/N, \ N \geq y.$$

This series diverges and so this produces an improper posterior distribution. Although our bounded prior is proper and so our posterior distribution is formally proper, this "almost improper" character is likely to lead to extreme sensitivity to prior assumptions. For example, a second run with $M = 15{,}000$ results in a

posterior mean of 3041 and median 1258. In fact we could show algebraically that the posterior mean increases as $M/\log(M)$; thus we can make it as big as we want by increasing M (proof as exercise).

We now consider Jeffreys' suggestion of a prior $p(N) \propto 1/N$, which is improper but can be constructed as follows if an upper bound, say 5000, is set.

```
N                ~ dcat(p.jeffreys[])
for (j in 1:5000) {
  reciprocal[j] <- 1/j
  p.jeffreys[j] <- reciprocal[j]/sum.recip
}
sum.recip       <- sum(reciprocal[])
```

The results show a posterior mean of 409 and median 197, which seems more reasonable — Jeffreys approximated the probability that there are more than 200 trams as $1/2$.

```
node mean  sd    MC error 2.5%  median 97.5%  start sample
N    408.7 600.4 4.99     102.0 197.0  2372.0 1001  10000
```

Suppose we now change the arbitrary upper bound to $M = 15{,}000$. Then the posterior mean becomes 520 and median 200. The median, but not the mean, is therefore robust to the prior. We could show that the conclusion about the median is robust to the arbitrary choice of upper bound M by proving that as M goes to infinity the posterior median tends to a fixed quantity (proof as exercise).

Finally, if a sensitivity analysis shows that the prior assumptions make a difference, then this finding should be welcomed. It means that the Bayesian approach has been worthwhile taking, and you will have to think properly about the prior and justify it. It will generally mean that, at a minimum, a weakly informative prior will need to be adopted.