L4: Jeffreys' Prior, Eliciting and Analysing Priors

1 Jeffreys' prior

1.1 Example of the problem of "uninformative" prior

Priors considered "uninformative" for a parameter, θ, may not be considered uninformative for transformations of the parameter, ψ=$g(\theta)$. For example, consider a Uniform(0,20) prior for a parameter θ. The problem with such uniform priors is that the *induced prior* for simple transformations of the parameter, e.g., ϕ=$\sqrt{\theta}$, will not be uniform. Given $\theta \sim$ Uniform(0,20), the distribution for ϕ is found by the change of variable theorem[1]:

$$\pi(\phi) = \frac{1}{20}\left|\frac{d\phi^2}{d\phi}\right| I(0 < \phi < \sqrt{20}) = \frac{1}{20}2\phi I(0 < \phi < \sqrt{20}) = \frac{\phi}{10} I(0 < \phi < \sqrt{20})$$

where $I(0 < \phi < \sqrt{20})$ is an indicator function that equals 1 when $0 < \phi < \sqrt{20}$ and 0 otherwise. This *induced prior* clearly is informative, as it linearly increases from 0 to $\sqrt{20}$.

1.2 Definition and calculation of Jeffreys prior

Jeffreys prior is an example of an *objective prior* which can be seen as a remedy to the just discussed problem of the induced priors resulting from transformations. It is a prior that is invariant to strictly monotonic (1-1 or bijective) transformations of the parameter, say ϕ =$g(\theta)$, where g is strictly monotonic.

Jeffreys prior is proportional to the square root of Fisher's Information, $I(\theta|y)$:

$$\text{Jeffreys prior } \pi_{JP}(\theta) \propto \sqrt{I(\theta|y)} \tag{1}$$

where

$$I(\theta|y) = \mathbb{E}\left[\left(\frac{d \log f(y|\theta)}{d\theta}\right)^2\right] \tag{2}$$

Note that under certain regularity conditions (e.g. that the differentiation operation can be moved inside the integral), Fisher Information can be calculated from the second derivative of the log likelihood:

$$I(\theta|y) = -\mathbb{E}\left[\frac{d^2 \log f(y|\theta)}{d\theta^2}\right] \tag{3}$$

which is often much easier to calculate than Eq'n 2.

[1] The change of variable theorem is a procedure for determining the pdf of a (continuous) random variable Y that is a strictly monotonic (1:1) transformation of another (continuous) random variable X, i.e., Y=$g(X)$. The pdf for Y:

$$p_Y(y) = p_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|$$

1.3 Jeffreys prior is invariant to 1:1 transformations

Remark. *Let $f(y|\theta)$ denote a probability density or mass function for a random variable y where θ is a scalar. Let $\phi = g(\theta)$ where g is a strictly monotonic (1:1 or bijective) transformation. If we specify a Jeffreys' prior for θ, namely, $\pi_{JP}(\theta) \propto \sqrt{I(\theta|y)}$, then the induced prior on ϕ, $\pi(\phi)$, is proportional to $\sqrt{I(\phi|y)}$. In other words, a 1:1 transformation of a parameter that has a Jeffreys prior yields a Jeffreys prior for the transformed parameter.*

Proof. This proof uses the chain rule, $\frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx}$, and the change of variable theorem.

Write the Fisher information for θ as follows:

$$\begin{aligned}
I(\theta|y) &= \mathbb{E}\left[\left(\frac{d\log f(x|\theta)}{d\theta}\right)^2\right] = \mathbb{E}\left[\left(\frac{d\log f(y|\phi))}{d\phi}\frac{d\phi}{d\theta}\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{d\log f(y|\phi))}{d\phi}\right)^2\right]\left|\frac{d\phi}{d\theta}\right|^2 = I(\phi|y)\left|\frac{d\phi}{d\theta}\right|^2
\end{aligned}$$

Thus the Jeffreys prior for θ can be written:

$$\pi_{JP}(\theta) \propto \sqrt{I(\theta|y)} = \sqrt{I(\phi|y)}\left|\frac{d\phi}{d\theta}\right|$$

Then the induced distribution for ϕ given this prior:

$$\pi(\phi) = \pi_{JP}(\theta)\left|\frac{d\theta}{d\phi}\right| \propto \sqrt{I(\phi|y)}\left|\frac{d\theta}{d\phi}\right|\left|\frac{d\phi}{d\theta}\right| = \sqrt{I(\phi|y)}$$

□

1.4 Example A. Binomial distribution

Suppose that the prevalence of Potato Virus Y in a population of aphids is an unknown parameter θ. A random sample of n aphids is taken (using a trap) and the number of aphids with the virus is x. Assuming independence between the aphids and that they all have the same probability of having the virus, $x \sim$ Binomial(n, θ). The Fisher information:

$$\begin{aligned}
I(\theta) &= -\mathbb{E}\left[\frac{d^2\left(\log\binom{n}{x} + x\log(\theta) + (n-x)\log(1-\theta)\right)}{d\theta^2}\right] = -\mathbb{E}\left[\frac{-x}{\theta^2} - \frac{n-x}{(1-\theta)^2}\right] \\
&= \frac{\mathbb{E}(x)}{\theta^2} + \frac{n-\mathbb{E}(x)}{(1-\theta)^2} = \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}
\end{aligned}$$

Thus the Jeffreys prior is

$$\pi(\theta) \propto \sqrt{\frac{1}{\theta(1-\theta)}} = \theta^{-1/2}(1-\theta)^{-1/2}$$

which is the kernel for a Beta(1/2,1/2) distribution.

Aside. Note that the mle for θ is $\hat{\theta} = x/n$ and the variance of $\hat{\theta}$ is $\frac{\theta(1-\theta)}{n}$, which equals $I(\theta)^{-1}$.

1.5 Example B. Exponential distribution

Let $x_1, \ldots, x_n$ be an iid sample from an exponential distribution with rate parameter λ, namely, $x_i \overset{iid}{\sim}$ Exponential(λ), $i = 1, \ldots, n$. For example, suppose that x_i is amount of time individual i waits in a queue at a bank during lunch hour until seeing a teller.

The Fisher information for a single random variable:

$$I(\lambda) = \mathbb{E}\left[\left(\frac{d \log f(x|\lambda)}{d\lambda}\right)^2\right] = \mathbb{E}\left[\left(\frac{d \log \lambda - \lambda x}{d\lambda}\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{\lambda} - x\right)^2\right] = \mathbb{E}\left[\frac{1}{\lambda^2} - 2\frac{x}{\lambda} + x^2\right]$$
$$= \frac{1}{\lambda^2} - \frac{2}{\lambda^2} + \frac{2}{\lambda^2} = \frac{1}{\lambda^2}$$

where we used the facts that $E(x)$=$1/\lambda$ and $E(x^2)$=$1/\lambda^2$ (use the moment generating function $\lambda/(\lambda - t)$). Alternatively, we could use the other expression for $I(\lambda)$:

$$I(\lambda) = -\mathbb{E}\left[\frac{d^2 \log f(x|\lambda)}{d\theta^2}\right] = -\mathbb{E}\left[\frac{-1}{\lambda^2}\right] = \frac{1}{\lambda^2}$$

Thus the Fisher information for the entire sample is then $I_n(\lambda) = \frac{n}{\lambda^2}$.

Then the Jeffreys prior:

$$\pi(\lambda) \propto \sqrt{I(\lambda)} = \frac{1}{\lambda}$$

Note that this is an *improper* prior as it does not integrate over the domain of λ, namely $(0, \infty)$. However this is a situation where the posterior for λ is proper:

$$p(\lambda|x_1, \ldots, x_n) \propto \pi(\lambda) f(x_1, \ldots, x_n|\lambda) = \frac{1}{\lambda}\lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i\right) = \lambda^{n-1} \exp\left(-\lambda \sum_{i=1}^n x_i\right)$$

which is the kernel for a Gamma($n, \sum_{i=1}^n x_i$) density function.

To demonstrate the invariance under a 1:1 transformation, reparameterize the exponential with θ=$g(\lambda) = \sqrt{\lambda}$, then

$$f(x|\theta) = \theta^2 \exp(-\theta^2 x)$$

and $g^{-1}(\theta) = \theta^2 = \lambda$. Given $\pi(\lambda) \propto 1/\lambda$, the induced prior for θ:

$$\pi_\theta(\theta) = \left|\frac{dg^{-1}(\theta)}{d\theta}\right| \pi_\lambda(g^{-1}(\theta)) = \left|\frac{d\theta^2}{d\theta}\right| \frac{1}{\theta^2} = \frac{2}{\theta}$$

Checking that this is indeed the Jeffreys prior for θ:

$$I(\theta) = -E\left[\frac{d^2 2\log(\theta) - \theta^2 x}{d\theta^2}\right] = \frac{4}{\theta^2}$$

Thus the Jeffreys prior is $\pi(\theta) \propto \sqrt{I(\theta)} = 2/\theta$.

1.6 Example C. Normal dist'n with known variance

The sampling distribution for $x_1, \ldots, x_n$ is Normal(μ, σ^2) where σ^2 is known. It can be shown that the Jeffreys prior for μ when σ^2 is known is

$$\pi(\mu) \propto 1, \ -\infty < \mu < \infty$$

This is an improper prior because $\int_{-\infty}^{\infty} 1 d\mu$ is not finite. However the posterior distribution is proper:

$$\pi(\mu|\boldsymbol{y}) \propto \exp\left(-\frac{(\mu-\bar{y})^2}{2\sigma^2/n}\right) * 1$$

namely the kernel for a Normal $\left(\bar{y}, \frac{\sigma^2}{n}\right)$.

1.7 Example D. Normal dist'n with known mean

The sampling distribution for $x_1, \ldots, x_n$ is Normal(μ, σ^2) where μ is known. The Jeffreys prior for σ^2 (given known μ) can be shown to be the following:

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}, \quad 0 < \sigma^2 < \infty \tag{4}$$

which is an improper prior because $\int_0^\infty \frac{1}{\sigma^2} d\sigma^2$ is not finite. The posterior distribution for σ^2,

$$p(\sigma^2|) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{z^2}{2\sigma^2}\right)\frac{1}{\sigma^2} = \left(\sigma^2\right)^{-(\frac{n}{2}+1)} \exp\left(-\frac{z^2}{2\sigma^2}\right) \tag{5}$$

where $z^2 = \sum_{i=1}^n (x_i-\mu)^2$. This is the kernel for $\Gamma^{-1}\left(\frac{n}{2}, \frac{z^2}{2}\right)$ so long as $n/2 > 0$ (obviously so), and $z^2 > 0$, which simply means that not all values in $\boldsymbol{y}$ are identical.

1.8 Jeffreys prior for a multivariate parameter vector.

To calculate Jeffreys prior for a multivariate parameter vector, θ, with p parameters, one calculates the score function, which is now the gradient of the log likelihood:

$$S(\theta) = \begin{bmatrix} \frac{d}{d\theta_1}\ln(f(x)) \\ \vdots \\ \frac{d}{d\theta_p}\ln(f(x)) \end{bmatrix} \tag{6}$$

and then calculates the Hessian of the log likelihood, namely, the matrix of the partial derivatives of the score function

$$H(\theta) = \begin{bmatrix} \frac{d}{d\theta_1}S(\theta)[1] & \frac{d}{d\theta_1}S(\theta)[2] & \ldots & \frac{d}{d\theta_1}S(\theta)[p] \\ \frac{d}{d\theta_2}S(\theta)[1] & \frac{d}{d\theta_2}S(\theta)[2] & \ldots & \frac{d}{d\theta_2}S(\theta)[p] \\ \vdots & \vdots & \vdots & \vdots \\ \frac{d}{d\theta_p}S(\theta)[1] & \frac{d}{d\theta_p}S(\theta)[2] & \ldots & \frac{d}{d\theta_p}S(\theta)[p] \end{bmatrix} = \begin{bmatrix} \frac{d^2}{d\theta_1^2}\ln(f(x)) & \frac{d^2}{d\theta_1 d\theta_2}\ln(f(x)) & \ldots & \frac{d^2}{d\theta_1 d\theta_p}\ln(f(x)) \\ \frac{d^2}{d\theta_2 d\theta_1}\ln(f(x)) & \frac{d^2}{d\theta_2^2 d\theta_2}\ln(f(x)) & \ldots & \frac{d^2}{d\theta_2 d\theta_p}\ln(f(x)) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{d^2}{d\theta_p d\theta_1}\ln(f(x)) & \frac{d^2}{d\theta_2 d\theta_p}\ln(f(x)) & \ldots & \frac{d^2}{d\theta_p^2}\ln(f(x)) \end{bmatrix} \tag{7}$$

Fisher information is

$$I(\theta|x) = -\mathbb{E}[H(\theta)] \tag{8}$$

Finally, the Jeffreys' prior for the vector of parameters is proportional to the square root of the determinant of the Fisher information matrix:

$$\pi(\theta) \propto \sqrt{\det(I(\theta|x))} \tag{9}$$

1.8.1 Example E. Normal(μ,σ^2) Jeffreys prior

To make the differentiation a little less awkward, the normal distribution is parameterized with θ=σ^2. Given $y_1, \ldots, y_n$ (iid) Normal(μ, θ) pdf can be written:

$$f(\boldsymbol{y}; \mu, \theta) = (2\pi\theta)^{-n/2} e^{-\frac{1}{2\theta}\sum_{i=1}^{n}(y_i-\mu)^2} \tag{10}$$

Then the log likelihood:

$$\log L(\mu, \theta) = l \propto -n\log(\theta) - \theta^{-1}\sum_{i=1}^{n}(y_i - \mu)^2 \tag{11}$$

The score vector:

$$\frac{\mathrm{d}l}{\mathrm{d}\mu} = 2\theta^{-1}\sum_{i=1}^{n}(y_i - \mu) = 2\theta^{-1}n(\overline{y} - \mu) \tag{12}$$

$$\frac{\mathrm{d}l}{\mathrm{d}\theta} = -n\theta^{-1} + \theta^{-2}\sum_{i=1}^{n}(y_i - \mu)^2 \tag{13}$$

The Hessian matrix (H) components:

$$\frac{\mathrm{d}^2 l}{\mathrm{d}\mu^2} = -2\theta^{-1}n \tag{14}$$

$$\frac{\mathrm{d}^2 l}{\mathrm{d}\mu\mathrm{d}\theta} = \frac{\mathrm{d}^2 l}{\mathrm{d}\theta\mathrm{d}\mu} = -2\theta^{-2}n(\overline{y} - \mu) \tag{15}$$

$$\frac{\mathrm{d}^2 l}{\mathrm{d}\theta^2} = n\theta^{-2} - 2\theta^{-3}\sum_{i=1}^{n}(y_i - \mu)^2 \tag{16}$$

Then the Fisher Information matrix

$$I(\mu, \theta) = -E[H] = \begin{bmatrix} \frac{2n}{\theta} & 0 \\ 0 & \frac{n}{\theta^2} \end{bmatrix} = \begin{bmatrix} \frac{2n}{\sigma^2} & 0 \\ 0 & \frac{n}{\sigma^4} \end{bmatrix} \tag{17}$$

Then the Jeffreys prior for (μ, σ^2):

$$\pi(\mu, \sigma^2) \propto \sqrt{|I(\mu, \sigma^2)|} = \sqrt{\frac{2n}{\sigma^2}\frac{n}{\sigma^4}} = \sqrt{\frac{2n^2}{\sigma^6}} \propto \frac{1}{(\sigma^2)^{\frac{3}{2}}} \tag{18}$$

The posterior can be shown to be the product of an inverse Chi-square (Gamma) and a normal (conditional on σ^2).

1.8.2 Example F. Normal with two independent Jeffreys priors.

Note: this example is Not using the Determinant as above.

If the Jeffreys prior for μ (given σ^2 is known) and the Jeffreys prior for σ^2 (given μ is known) are treated as independent priors, the Jeffreys prior is

$$\pi(\mu, \sigma^2) \propto 1 * \frac{1}{\sigma^2}, \ -\infty < \mu < \infty, \ 0 < \sigma^2 < \infty$$

then it turns out that the posterior marginal distribution for σ^2 is

$$\sigma^2|\boldsymbol{y} \sim \Gamma^{-1}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$$

where $s^2 = (1/(n-1))\sum_{i=1}^{n}(y_i - \bar{y})^2$, the usual sample variance. And the posterior marginal distribution for μ is a student's t distribution with mean $\bar{y}$, scale parameter s^2/n, and $n-1$ degrees of freedom.

Referring to the 2x4 timber example from Lecture 3 notes, $\bar{y}$=3.5283, s^2= 0.0122209, with 9 degrees of freedom. Thus the posterior expected value of μ is 3.5283. A 95% credible interval for μ can be found by calculating the 2.5 and 97.5 percentiles of the standard $t_{9\ df}$ distribution, multiplying them by the square root of the scale parameter and then adding the mean:

```
qt(c(0.025,0.975),df=9)*sqrt(0.0122209/10) + 3.5283
 3.449219  3.607381
```

Thus 95% credible interval for μ of (3.45, 3.61).

For σ^2, the 95% credible interval is calculated for the Gamma(9/2, 9*0.0122209/2) distribution and then inverting the results:

```
gam.bds <- qgamma(c(0.025,0.975),shape=9/2,rate=9*0.0122209/2)
inv.gam.bds <- sort(1/gam.bds)
inv.gam.bds
0.005781919 0.040730458
```

Thus a 95% credible interval for σ^2 of (0.0058, 0.0407).

2 Reference Priors

As mentioned in LN 2, the Jeffreys' prior is one of several objective priors, namely procedures for selecting priors which will yield the same prior for anyone who uses the procedure. Reich and Ghosh (2019) discuss four other objective priors that are at least worth knowing about and I recommend reading the approximately two pages of discussion. Admittedly understanding the general concepts behind the procedures and being able to implement the procedures can have quite different degrees of difficulty, with the latter generally more difficult than the former. Here we will just examine one other type of objective prior, the Reference Prior, denoted $\pi_{RP}(\theta)$.

2.1 KL divergence

Before introducing Reference Priors, the notion of Kullback-Leibler (KL) divergence is introduced. KL divergence is a measure of the difference between two pmfs or two pdfs. A KL divergence value of 0 means that the two distributions are identical.

For the continuous case let f and g denote two pdfs. The KL divergence is defined "conditional" on one of the two distributions, here denoted $KL(f, g)$ or $KL(g, f)$ where $KL(f, g) \neq KL(g, f)$, except when f and g are identical (almost everywhere). More exactly, $KL(f, g)$ is the expected value of $\log(f(x)/g(x)$ assuming that f is "true"—more accurately stated that the expectation is with respect to $f(x)$, and $KL(g, f)$ is the reverse:

$$KL(f, g) = \int \log\left[\frac{f(x)}{g(x)}\right] f(x)dx = \int \log(f(x))f(x)dx - \int \log(g(x))f(x)dx = E_f\left[\log(f(X)\right] - E_f\left[\log(g(X)\right] \tag{19}$$

and

$$KL(g, f) = \int \log\left[\frac{g(x)}{f(x)}\right] g(x)dx = E_g\left[\log(g(x))\right] - E_g\left[\log(f(x)\right] \tag{20}$$

Notes: (1) if $g(x) = f(x)$, then $\log\left[\frac{f(x)}{g(x)}\right] = \log(1) = 0$; and (2) $KL(f, g) \geq 0$.

KL divergence examples. As a simple example consider a discrete valued random variable with values 0, 1, or 2. Let $f(x)$ be the Binomial(2,p=0.2) pmf with probabilities 0.64, 0.32, and 0.04 for X=0, 1, and 2, respectively. Let $g(x)$ be the discrete uniform where $g(0) = g(1) = g(2) = 1/3$. Then

$$KL(f, g) = \sum_{x=0}^{2} f(x)\log\left[\frac{f(x)}{g(x)}\right] = 0.64\log(0.64/0.33) + 0.32\log(0.32/0.33) + 0.04\log(0.04/0.33) = 0.3196145$$

$$KL(g, f) = \sum_{x=0}^{2} g(x)\log\left[\frac{g(x)}{f(x)}\right] = 0.33\log(0.33/0.64) + 0.33\log(0.33/0.32) + 0.33\log(0.33/0.04) = 0.5029201$$

For another example let $g(x)$ be a Binomial(2,p=0.25). Then $KL(f, g)$= 0.01400421 and $KL(g, f)$ = 0.01476399, thus the KL divergence measures are both close to 0 (and close to each other).

2.2 Reference prior

Given data $\mathbf{y}$, the KL divergence between the prior and posterior (with respect to the posterior) is

$$KL(p(\theta|\mathbf{y}), \pi(\theta)) = \int p(\theta|\mathbf{y})\log\left[\frac{p(\theta|y)}{\pi(\theta)}\right] d\theta \tag{21}$$

The key idea of a RP is that the KL divergence between the posterior and the prior should be as large as possible, thus implying that the data are dominating the prior.

However, this measure is conditional on the data, $\mathbf{y}$, which does not help for determining a prior. Thus to remove the conditioning on the data, the data are integrated out, and $\pi_{RP}(\theta)$ is the probability distribution (pmf or pdf) that maximizes the following:

$$E_{\mathbf{y}}\left[KL(p(\theta|\mathbf{y}), \pi(\theta))\right] = \int \left[KL(p(\theta|\mathbf{y}), \pi(\theta))\right] m(\mathbf{y})d\mathbf{y} \tag{22}$$

where $m(\mathbf{y})$ is the marginal distribution for the data.

While this approach is conceptually attractive, it can be technically challenging as one is trying to find an entire probability distribution $\pi(\theta)$ that maximizes eq'n (22), noting that determining $m(\mathbf{y})$ involves integration as well.

3 Eliciting Informative Priors

Often the scientist or subject-matter specialist will have a definite opinion as to what the range of parameter values should be. For example, an experienced heart surgeon who has done 1000s of coronary by-pass surgeries on a variety of patients will have a definite opinion on post-surgery survival probability.

How does one translate that prior knowledge into a prior probability distribution?

- First, think about whether it's even feasible or are there too many parameters, or is the underlying sampling model fairly complex, e.g., a hierarchical model. For example, suppose the sampling model is a multiple regression with 4 covariates, thus these five parameters, β_0, β_1, β_2, β_3, and β_4, and the variance parameter. How might the expert's knowledge be translated into prior distributions for these 5 parameters? The expert might have an opinion about the signs, positive or negative, of each coefficient, and perhaps the relative importance of each covariate; e.g., dealing with standardized covariates, then the effects of x_1 and x_2 are thought to be positive but the effect of x_1 may be twice as large as x_2.

- Single parameter case. This is generally the most feasible situation. If the expert is not familiar with probability distributions, the statistician may need to work with the expert to arrive at a prior and can help by asking questions about the parameter without using statistics jargon.

  For example, instead of asking for the median, ask "For what value of the parameter do you think that it's equally likely that values are either below or above it?"

  "What do you think the range of values might be?"

  " What do you think the relative variation around an average value be, for example, if your best guess for θ is 15, is your uncertainty within ± 10% of that value (± 1.5), or 20% (± 3.0)?" Thus potentially getting a measure of the coefficient of variation, $CV = \sigma/\mu$.

  Given a mean value, and a range, standard deviation, or CV, and assuming a particular standard probability distribution might suffice, rough estimates of hyperparameters for the prior might be calculated. This is the "moment matching" idea that we've examined previously.

  For example, with the above example of the surgeon, the surgeon was thinking that θ would be 0.7 on average. Further questioning about the surgeon's uncertainty led to a determination that a CV of 0.1 would be appropriate. Using a Beta distribution for the prior, the mean is $\alpha/(\alpha+\beta)$ and the variance is $(\alpha\beta)/[(\alpha+\beta)^2(\alpha+\beta+1)]$, some algebra yields Beta(29.3, 12.6).

- Discrete histogram priors. Another simple way to elicit priors is to partition parameter values into non-overlapping bins, and have the expert present relative weights for each bin. For example, θ is grouped into three bins, [0,10), [10,25), [25,30], and the expert gives relative weights of 0.2, 0.5, and 0.3. A proper histogram, pdf, is constructed using the result that bin area = height × width, where bin area corresponds to probability or weight. For the bin [0,10), area=0.2, width=10, thus height equals area/width = 0.2/10 = 0.02. For [10,25) height is 0.5/15 = 0.033, and for [25,30] height is 0.3/5 = 0.06.

3.1 References

- "The elicitation of prior distributions", Chaloner, 1996, in *Bayesian Biostatistics*, eds., Berry and Stangl.

- *Uncertain Judgements: Eliciting Experts' Probabilities*, O'Hagan, et al. 2006. This book is available online as a pdf via the University Library.

4 Sensitivity analysis of priors

Using the term *sensitivity analysis* loosely here, we mean an examination of the effects of different priors on the posterior. For example, the comparison of the posterior distribution for the probability of survival after by-pass surgery for the surgeon's prior and the medical student's prior is a sensitivity analysis.

Another loosely put phrase, if the posterior distributions for different priors look much "the same", e.g., have similar means and variances, then then one might say that the results are *robust* to the priors.

With large enough samples, unless the prior is particularly concentrated over a narrow range of possible values, sometimes called a pig-headed prior, the posterior will look much the same for a wide range of priors as the data (the likelihood) are *dominating* the prior.

Problematic issues

- Practical issue: if 100s of parameters, then tedious at least, to carry out a sensitivity analysis for all the priors.
- Generalized linear models where data come from an exponential family distribution with parameter θ and covariate(s) x, say $\mathcal{F}$, and, $g(\theta, x)$, the "link" function, is a linear model:

  $$\begin{aligned} y|\theta, x &\sim \mathcal{F}(\theta, x) \\ g(\theta, x) &= \beta_0 + \beta_1 x \end{aligned}$$

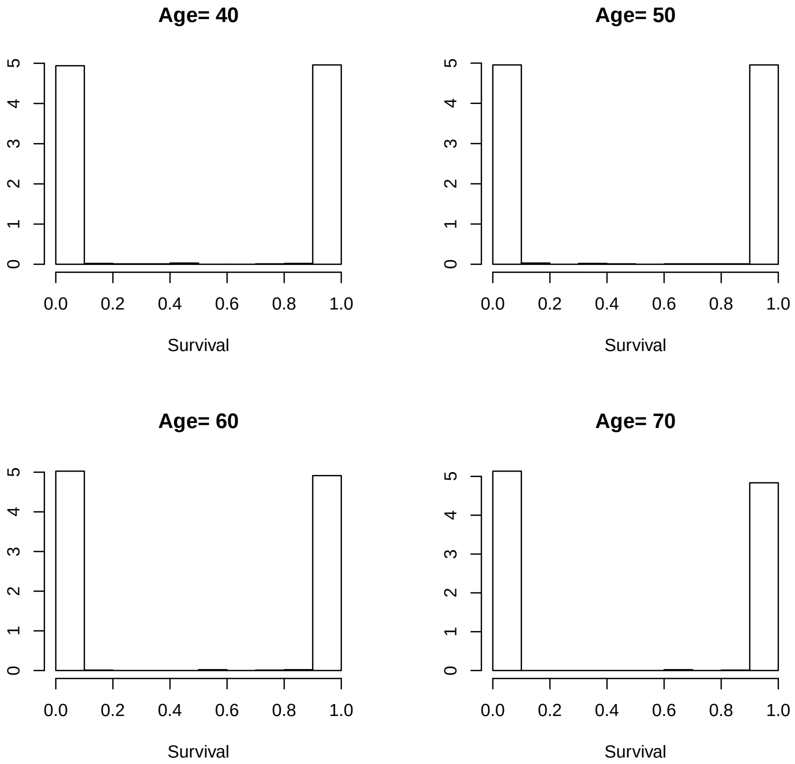
  Apparently uninformative priors in the link function may *induce* quite informative priors at a lower level. For example, a logistic regression for the number of patients surviving heart bypass surgery where the probabilities differ with age:

  $$\begin{aligned} y_i|Age_i &\sim \text{Bernoulli}(\theta(Age_i)) \\ &\text{where } \theta(Age) = \frac{\exp(\beta_0 + \beta_1 Age)}{1 + \exp(\beta_0 + \beta_1 Age)} \\ &\text{equivalently } \ln\left(\frac{\theta(Age)}{1 - \theta(Age)}\right) = \beta_0 + \beta_1 Age \end{aligned}$$

  A seemingly innocuous prior for both β_0 and β_1 is Normal(μ=0,σ^2=5^2). Suppose that the ages range from 40 to 70. Figure 1 shows the results of simulating from the priors for β_0 and β_1 on the induced priors for survival for four different ages. Note how the probabilities are massed near 0 and 1. Such simulation exercises can be quite valuable for detecting such effects.

October 10, 2020

Figure 1: Induced prior for survival probability (θ) as a function of age given Normal(0,5) priors for logit transformation, $\ln(\theta/(1-\theta)) = \beta_0 + \beta_1 Age$.


Age= 40
Age= 50
Age= 60
Age= 70
Survival