MAS3301 Bayesian Statistics

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 2, 2008-9

7 Numerical Methods for More Than One Parameter

7.1 Introduction

It is often necessary to use numerical methods to do the necessary integrations for computing posterior distributions and summaries. We have already seen, in section 6, the use of simple numerical integration when we have just one unknown parameter. Similar methods can be used when we have more than one. We will look at this first in the case of two unknown parameters.

If we have two unknown parameters $\theta_1,\ \theta_2$ then we often need to create a two-dimensional grid of values, containing every combination of $\theta_{1,1},\ldots,\theta_{1,m_1}$ and $\theta_{2,1},\ldots,\theta_{2,m_2}$, where $\theta_{j,1},\ldots,\theta_{j,m_j}$ are a set of, usually equally spaced, values of θ_j. We therefore have $m_1 m_2$ points and two step sizes, $\delta\theta_1,\ \delta\theta_2$. Figure 14 shows such a grid diagramatically. Instead of a collection of two-dimensional rectangular columns standing on a one-dimensional line, we now have a collection of three-dimensional rectangular columns standing on a two-dimensional plane. The contours in figure 14 represent the function being integrated. The small circles represent the points at which the function is evaluated. The dashed lines represent the boundaries of the columns. Of course we would really have many more function evaluations placed much more closely together. Notice that some of the function evaluations are in regions where the value of the function is very small. It is inefficient to waste too many function evaluations in this way and some more sophisticated methods avoid doing this.

The approximate integral becomes

$$\int\int h(\theta_1,\theta_2)\ d\theta_1\ d\theta_2 \approx \sum_{j=1}^{m_1}\sum_{k=1}^{m_2} h(\theta_{1,j},\theta_{2,k})\delta\theta_1\delta\theta_2.$$

We can extend this to three or more dimensions but it becomes impractical when the number of dimensions is large. If we use a 100×100 grid in two dimensions this gives 10^4 function evaluations. If we use a $100\times100\times100$ grid in three dimensions this requires 10^6 evaluations and so on. Clearly the number of evaluations becomes prohibitively large quite quickly as the number of dimensions increases. In such cases we would usually use Markov chain Monte Carlo methods which are beyond the scope of this module.

It is sometimes possible to reduce the dimension of the numerical integral by integrating analytically with respect to one unknown.

7.2 Example: The Weibull distribution

7.2.1 Model

The *Weibull distribution* is often used as a distribution for *lifetimes*. We might be interested, for example, in the lengths of time that a machine or component runs before it fails, or the survival time of a patient after a serious operation. A number of different families of distributions are used for such lifetime variables. Of course they are all continuous distributions and only give positive probability density to positive values of the lifetime. The Weibull distribution is an important distribution of this type. We can think of it as a generalisation of the exponential distribution. The distribution function of an exponential distribution is $F(t) = 1 - \exp(-\lambda t)$. The distribution function of a Weibull distribution is

$$F(t) = 1 - \exp\left(-\lambda t^\alpha\right) \qquad (t \geq 0) \tag{5}$$

where the extra parameter $\alpha > 0$ is called a *shape parameter*. It is often convenient to write $\lambda = \rho^\alpha$ and then

$$F(t) = 1 - \exp\left(-[\rho t]^\alpha\right) \qquad (t \geq 0) \tag{6}$$

and $\rho > 0$ is a *scale parameter*.

Differentiating (6) with respect to t, we obtain the pdf

$$f(t) = \alpha\rho(\rho t)^{\alpha-1}\exp\{-(\rho t)^\alpha\} \tag{7}$$

for $0 \leq t < \infty$.

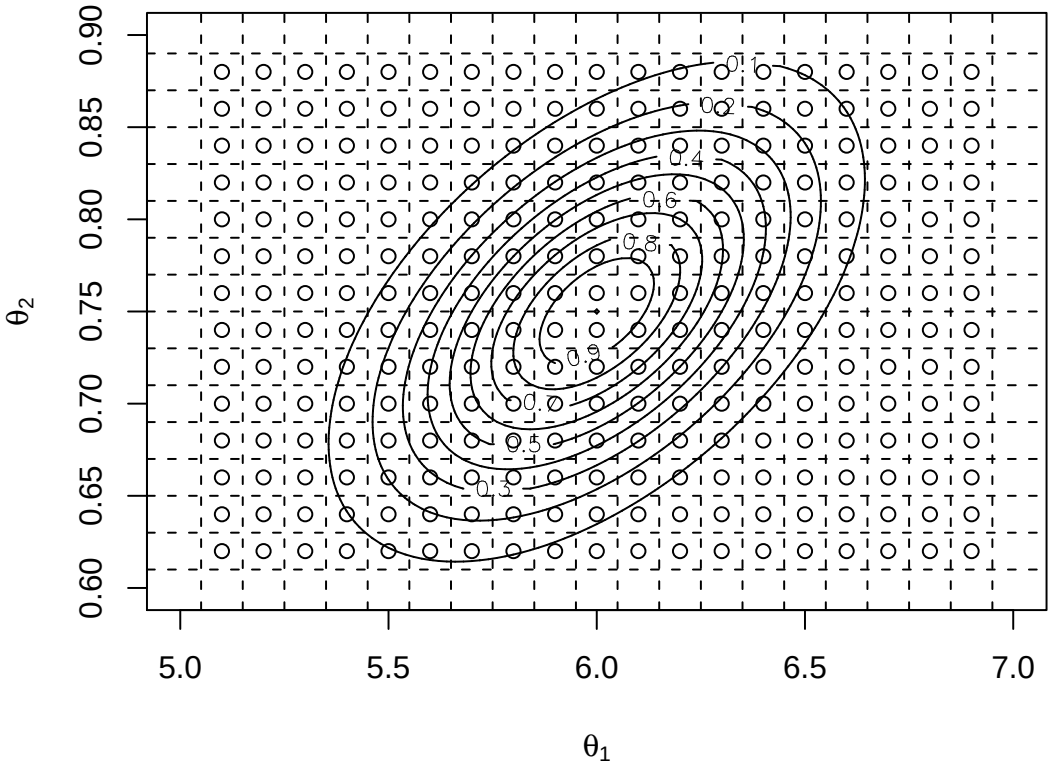

Figure 14: Numerical integration in two dimensions.

67	313	1391	630	627	573	2093	28	492	482
206	1166	165	1088	496	313	437	815	436	17
32	131	340	939	247	1859	57	132	813	254
950	1615	463	258	2285	672	506	50	637	246
178	431	306	662	33	254	858	187	344	545

Table 4: Data for Weibull example.

If we use $\alpha,\ \lambda$ instead of $\alpha,\ \rho$ as the parameters, as in (5), then the pdf is

$$f(t) = \alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha}). \tag{8}$$

7.2.2 Evaluating the posterior distribution

Suppose that we work in terms of the $\alpha,\ \rho$ parameters of (7) and that we have n observations $t_1, \ldots, t_n$. Suppose that our prior density for α and ρ is $f^{(0)}_{\alpha,\rho}(\alpha, \rho)$. The likelihood is

$$L(\alpha, \rho) = \alpha^n \rho^{n\alpha} \left(\prod_{i=1}^{n} t_i\right)^{\alpha-1} \exp\left\{-\rho^{\alpha}\sum_{i=1}^{n} t_i^{\alpha}\right\}.$$

The posterior pdf is

$$f^{(1)}_{\alpha,\rho}(\alpha, \rho) \propto h_{\alpha,\rho}(\alpha, \rho) = f^{(0)}_{\alpha,\rho}(\alpha, \rho)L(\alpha, \rho).$$

To complete the evaluation of the posterior pdf we find

$$C = \int_0^{\infty}\int_0^{\infty} h_{\alpha,\rho}(\alpha, \rho)\ d\alpha\ d\rho$$

numerically and then

$$f^{(1)}_{\alpha,\rho}(\alpha, \rho) = h_{\alpha,\rho}(\alpha, \rho)/C.$$

Suppose, for example, that we give α and ρ independent gamma prior distributions so that

$$f^{(0)}_{\alpha,\rho}(\alpha, \rho) \propto \alpha^{a_\alpha - 1}e^{-b_\alpha \alpha}\rho^{a_\rho - 1}e^{-b_\rho \rho}.$$

Then the posterior pdf is proportional to

$$h_{\alpha,\rho}(\alpha, \rho) = \alpha^{n+a_\alpha-1}\rho^{n\alpha+a_\rho-1}\left(\prod_{i=1}^{n} t_i\right)^{\alpha-1} \exp\left\{-\left[b_\alpha\alpha + b_\rho\rho + \rho^{\alpha}\sum_{i=1}^{n} t_i^{\alpha}\right]\right\}.$$

Figure 6 shows the posterior density of α and ρ when $n = 50,\ a_\alpha = 1,\ b_\alpha = 1,\ a_\rho = 3,\ b_\rho = 1000$ and the data are as given in table 4. Figure 7 shows the same thing as a perspective plot except that, to make the axes more readable, ρ has been replaced with $R = 1000\rho$.

To find, for example, the posterior mean of ρ we evaluate

$$\int_0^{\infty}\int_0^{\infty} \rho\ f^{(1)}_{\alpha,\rho}(\alpha, \rho)\ d\alpha\ d\rho = C^{-1}\int_0^{\infty}\int_0^{\infty} \rho\ h_{\alpha,\rho}(\alpha, \rho)\ d\alpha\ d\rho.$$

To find a 95 % hpd region for α, ρ we can either choose a value k and evaluate $\int\int f^{(1)}_{\alpha,\rho}(\alpha,\rho)\, d\alpha\, d\rho$ over all points in a grid for which $f^{(1)}_{\alpha,\rho}(\alpha,\rho) > k$ then adjust k and repeat until the value of 0.95 is obtained or rank all of the points in our grid in decreasing order of $f^{(1)}_{\alpha,\rho}(\alpha,\rho)$ and cumulatively integrate over them until 0.95 is reached.

To find the marginal pdf for α we evaluate

$$\int_0^\infty f^{(1)}_{\alpha,\rho}(\alpha,\rho)\ d\rho.$$

7.3 Transformations

7.3.1 Theory

It has probably become apparent by now that sometimes it may be helpful to use a transformation of the parameters. For example, sometimes a posterior distribution where we need to use numerical integration might have an awkward shape which makes placing a suitable and efficient rectangular grid difficult.

In section 5.4 we saw how to change the pdf when we transform a single random variable. Sometimes, of course, we need a more general method for transforming between one set of parameters and another. Let $\underline{\theta}$ and $\underline{\phi}$ be two alternative sets of parameters where there is a 1 - 1 relationship between values of $\underline{\theta}$ and values of $\underline{\phi}$, and therefore each contains the same number of parameters. (There could appear to be more parameters in $\underline{\theta}$ than in $\underline{\phi}$, for example, but, in that case, there would have to be constraints on the values of $\underline{\theta}$ so that there was the same *effective* number of parameters in $\underline{\theta}$ and $\underline{\phi}$). Let $\underline{\theta} = (\theta_1, \ldots, \theta_k)^T$ and $\underline{\phi} = (\phi_1, \ldots, \phi_k)^T$. Suppose also that we can write, for each i,

$$\phi_i = g_i(\theta_1, \ldots, \theta_k)$$

where g is a differentiable function. Then, if the density of $\underline{\theta}$ is $f_{\underline{\theta}}(\underline{\theta})$ and the density of $\underline{\phi}$ is $f_{\underline{\phi}}(\underline{\phi})$,

$$f_{\underline{\theta}}(\underline{\theta}) = f_{\underline{\phi}}(\underline{\phi})|J|$$

where J is the *Jacobian determinant*, often just called "the Jacobian,"

$$\begin{vmatrix} \frac{\partial\phi_1}{\partial\theta_1} & \frac{\partial\phi_1}{\partial\theta_2} & \cdots & \frac{\partial\phi_1}{\partial\theta_k} \\ \frac{\partial\phi_2}{\partial\theta_1} & \frac{\partial\phi_2}{\partial\theta_2} & \cdots & \frac{\partial\phi_2}{\partial\theta_k} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial\phi_k}{\partial\theta_1} & \frac{\partial\phi_k}{\partial\theta_2} & \cdots & \frac{\partial\phi_k}{\partial\theta_k} \end{vmatrix}$$

and $|J|$ is its modulus.

For example, we could transform the $(0, \infty)$ ranges of the parameters α, ρ of a Weibull distribution to $(0, 1)$ by using

$$\beta = \frac{\alpha}{\alpha + 1}, \qquad \gamma = \frac{\rho}{\rho + 1}.$$

The Jacobian is

$$J = \begin{vmatrix} \frac{\partial\beta}{\partial\alpha} & \frac{\partial\beta}{\partial\rho} \\ \frac{\partial\gamma}{\partial\alpha} & \frac{\partial\gamma}{\partial\rho} \end{vmatrix} = (\alpha + 1)^{-2}(\rho + 1)^{-2}.$$

Suppose that the joint posterior density of α and ρ is proportional to $h_{\alpha,\rho}(\alpha, \rho)$. So we define

$$h_{\beta,\gamma}(\beta, \gamma) = (\alpha + 1)^2(\rho + 1)^2 h_{\alpha,\rho}(\alpha, \rho),$$

where

$$\alpha = \frac{\beta}{1 - \beta}, \qquad \rho = \frac{\gamma}{1 - \gamma}$$

so

$$h_{\beta,\gamma}(\beta,\gamma) = (1-\beta)^{-2}(1-\rho)^{-2}h_{\alpha,\rho}\left(\frac{\beta}{1-\beta},\frac{\gamma}{1-\gamma}\right).$$

Then let

$$C = \int_0^1 \int_0^1 h_{\beta,\gamma}(\beta,\gamma)\ d\beta\ d\gamma.$$

The posterior mean of ρ is then

$$C^{-1}\int_0^1 \int_0^1 \frac{\gamma}{1-\gamma}h_{\beta,\gamma}(\beta,\gamma)\ d\beta\ d\gamma.$$

A hpd region for α, ρ can then be found by integrating $C^{-1}h_{\beta,\gamma}(\beta,\gamma)$ with respect to β, γ over the points with the greatest values of

$$h_{\alpha,\rho}\left(\frac{\beta}{1-\beta},\frac{\gamma}{1-\gamma}\right) = h_{\alpha,\rho}(\alpha,\rho).$$

7.3.2 Example: A clinical trial

The Anturane Reinfarction Trial Research Group (1980) reported a clinical trial on the use of the drug sulfinpyrazone in patients who had suffered myocardial infarctions ("heart attacks"). The idea was to see whether the drug had an effect on the number dying. Patients in one group were given the drug while patients in another group were given a "placebo," that is an inactive substitute. The following table gives the number of all "analysable" deaths up to 24 months after the myocardial infarction and the total number of eligible patients who were not withdrawn and did not suffer a "non-analysable" death during the study.

	Deaths	Total
Group 1 (Sulfinpyrazone)	44	560
Group 2 (Placebo)	62	540

We can represent this situation by saying that there are two groups, containing n_1 and n_2 patients, and two parameters, $\theta_1,\ \theta_2$, such that, given these parameters, the distribution of the number of deaths X_j in Group j is binomial$(n_j,\ \theta_j)$.

Now we could give θ_j a beta prior distribution but it seems reasonable that our prior beliefs would be such that θ_1 and θ_2 would not be independent. There are various ways in which we could represent this. One of these is as follows. We transform from the $(0,1)$ scale of $\theta_1,\ \theta_2$ to a $(-\infty,\infty)$ scale and then give the new parameters, $\eta_1,\ \eta_2$, a bivariate normal distribution (see section 5.3). We can use a transformation where $\theta_j = F(\eta_j)$ and $F(x)$ is the distribution function of a continuous distribution on $(-\infty,\ \infty)$, usually one which is symmetric about $x = 0$. One possibility is to use the standard normal distribution function $\Phi(x)$ so that $\theta_j = \Phi(\eta_j)$. We write $\eta_j = \Phi^{-1}(\theta_j)$ where this function, $\Phi^{-1}(x)$, the inverse of the standard normal distribution function, is sometimes called the *probit* function. If we use this transformation then it is easily seen that

$$f_\theta(\theta_1,\theta_2) = f_\eta(\eta_1,\eta_2)/|J|,$$

where $f_\theta(\theta_1,\theta_2)$ is the joint density of $\theta_1,\ \theta_2$, $f_\eta(\eta_1,\ \eta_2)$ is the joint density of $\eta_1,\ \eta_2$ and

$$|J| = |\begin{vmatrix} \frac{\partial\theta_1}{\partial\eta_1} & \frac{\partial\theta_1}{\partial\eta_2} \\ \frac{\partial\theta_2}{\partial\eta_1} & \frac{\partial\theta_2}{\partial\eta_2} \end{vmatrix}| = \phi(\eta_1)\phi(\eta_2),$$

where $\phi(x)$ is the standard normal pdf.

Suppose that, from past experience, we can give a 95% symmetric prior interval for θ_2 (placebo) as $0.05 < \theta_2 < 0.20$. (This is actually quite a wide interval considering that there may be a lot of past experience of such patients). This converts to a 95% interval of $-1.645 < \eta_2 < -0.842$. For example, in R, we can use

```
> qnorm(0.05,0,1)
[1] -1.644854
```

If we give η_2 a normal prior distribution then we require the mean to be $\mu_2 = ([-1.645] + [-0.842])/2 \approx -1.24$ and the standard deviation to be $\sigma_2 = ([-0.842] - [-1.645])/(2 \times 1.96) \approx 0.21$, since a symmetric 95% normal interval is the mean plus or minus 1.96 standard deviations. Let us use the same mean for a normal prior distribution for η_1 (sulfinpyrazone) so that we have equal prior probabilities for an increase and a decrease in death rate when the treatment is given. However it seems reasonable that we would be less certain of the death rate given the treatment so we increase the prior standard deviation to $\sigma_1 = 2\sigma_2 = 0.42$. This implies a 95% interval $-2.06 < \eta_1 < -0.42$ which, in turn, implies $0.02 < \theta_1 < 0.34$. (This is a wide interval so we are really not supplying much prior information).

We also need to choose a covariance or correlation between η_1 and η_2. At this point we will not discuss in detail how to do this except to say that, if we choose the correlation to be r, then the conditional variance of one of $\eta_1,\ \eta_2$ given the other will be $100r^2\%$ of the marginal variance. For example, if we choose $r = 0.7$, then the variance of one is roughly halved by learning the value of the other. Suppose that we choose this value. Then the covariance between η_1 and η_2 is $0.7 \times 0.21 \times 0.42 = 0.0617$.

In evaluating the joint prior density of $\eta_1,\ \eta_2$, we can make use of the fact, which is easily confirmed, that, if $\delta_j = (\eta_j - \mu_j)/\sigma_j$ and $r = \text{covar}(\eta_1, \eta_2)/(\sigma_1\sigma_2)$, then the joint density is proportional to

$$\exp\left\{-\frac{1}{2(1-r^2)}(\delta_1^2 + \delta_2^2 - 2r\delta_1\delta_2)\right\}.$$

Figure 15 shows a R function to evaluate the posterior density. Figure 16 shows the resulting posterior density. The dashed line is the line $\theta_1 = \theta_2$. We see that most of the probability lies on the side where $\theta_2 > \theta_1$ which suggests that the death rate is probably greater with the placebo than with sulfinpyrazone, which, of course, suggests that sulfinpyrazone has a beneficial effect.

To investigate further what the posterior tells us about the effect of sulfinpyrazone, we can calculate the posterior probability that $\theta_1 < \theta_2$. This is done by integrating the joint posterior density over the region where $\theta_1 < \theta_2$. This calculation is included in the function shown in figure 15. The calculated probability is 0.972. We can also find the posterior density of the *relative risk*, θ_1/θ_2, or the *log relative risk*, $\log(\theta_1/\theta_2)$. Let γ be the log relative risk. We can modify the function in figure 15 so that it uses a grid of γ and θ_2 values, evaluates the joint posterior density of γ and θ_2 and then integrates out θ_2. Of course we need to transform between $\theta_1,\ \theta_2$ and $\gamma,\ \theta_2$ where the densities are related by

$$f_{\theta_1,\theta_2}(\theta_1, \theta_2) = f_{\gamma,\theta_2}(\gamma, \theta_2)|J|$$

and $J = \theta_1^{-1}$ is the appropriate Jacobian. Figure 17 shows the prior and posterior densities of the log relative risk, γ. Values of γ less than zero correspond to a smaller death rate with sulfinpyrazone than with the placebo. Notice that the prior density is not quite symmetric about zero. It is symmetric on the η scale but not on the γ scale. The prior median is zero, however.

There are other methods available to deal with problems of this sort, some involving approximations and fairly simple calculations.

8 Conjugate Priors I: The beta distribution

8.1 Introduction

We have already come across the idea of a conjugate prior distribution in Lecture 4. Now we will look at this idea in more detail.

Suppose we wish to learn about a parameter θ. Suppose our prior distribution for θ has pdf $f_0(\theta)$. Suppose we observe data y and the likelihood function is $L(\theta;\ y)$. Now our posterior pdf for θ is

$$f_1(\theta) \propto f_0(\theta)L(\theta;\ y).$$

```
function(theta1,theta2,n,x,prior)
{# Evaluates posterior density for probit example.
 # prior is mean1, mean2, sd1, sd2, correlation
 n1<-length(theta1)
 n2<-length(theta2)
 step1<-theta1[2]-theta1[1]
 step2<-theta2[2]-theta2[1]
 theta1<-matrix(theta1,nrow=n1,ncol=n2)
 theta2<-matrix(theta2,nrow=n1,ncol=n2,byrow=T)
 eta1<-qnorm(theta1,0,1)
 eta2<-qnorm(theta2,0,1)
 delta1<-(eta1-prior[1])/prior[3]
 delta2<-(eta2-prior[2])/prior[4]
 r<-prior[5]
 d<-1-r^2
 logprior<- -(delta1^2 + delta2^2 - 2*r*delta1*delta2)/(2*d)
 J<-dnorm(eta1,0,1)*dnorm(eta2,0,1)
 logprior<-logprior-log(J)
 loglik<-x[1]*log(theta1)+(n[1]-x[1])*log(1-theta1)+x[2]*log(theta2)+(n[2]-x[2])*log(1-theta2)
 logpos<-logprior+loglik
 logpos<-logpos-max(logpos)
 posterior<-exp(logpos)
 int<-sum(posterior)*step1*step2
 posterior<-posterior/int
 prob<-sum(posterior*(theta1<theta2))*step1*step2
 ans<-list(density=posterior,prob=prob)
 ans
 }
```

Figure 15: R function for probit example (7.3.2).

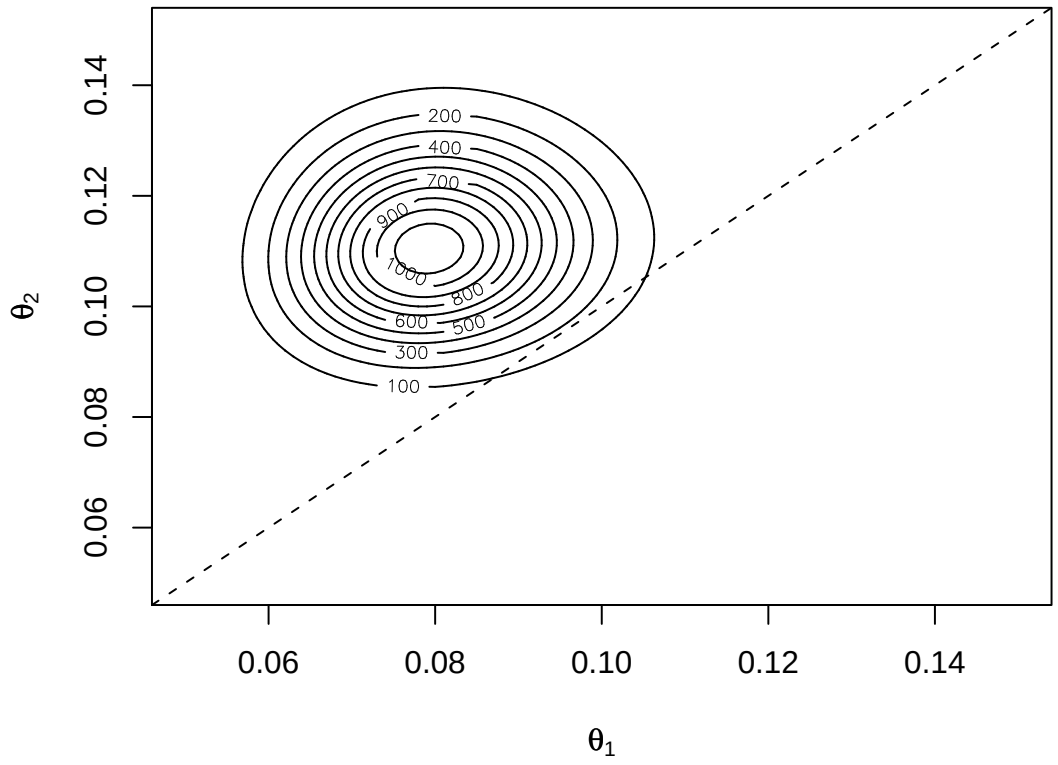

Figure 16: Posterior density of θ_1 and θ_2 in probit example (7.3.2). The dashed line is $\theta_1 = \theta_2$.

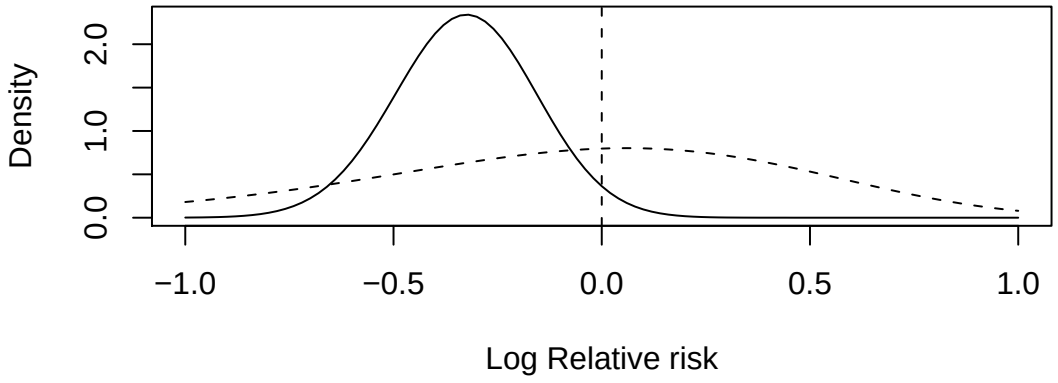

Figure 17: Posterior density (solid) and prior density (dashes) of log relative risk in probit example (7.3.2).

Conjugate prior distributions A *conjugate* prior distribution is one where the form of the posterior distribution is the same as the form of the prior distribution. That is $f_0(\theta)$ and $f_1(\theta)$ are pdfs belonging to the same family of (e.g. they are both normal or they are both gamma). Clearly the form of the conjugate prior depends on the form of the likelihood. So there is a different conjugate distribution for a different data (or "sampling") distribution. In some cases there is no conjugate prior distribution.

We are not, of course, compelled to use a conjugate prior distribution. Our prior beliefs simply might not be of the appropriate form. However, very often, a conjugate prior will be able to represent our beliefs closely enough and the calculations then become much easier. When we feel that we can not use a conjugate prior, even though one exists, we may be able to use a *mixture* of conjugate priors to represent our beliefs. This topic will be discussed in a later lecture.

8.2 Beta distribution

In this lecture we will look at the *beta* distribution as a conjugate prior distribution.

If $\theta \sim \text{beta}(a, b)$ then it has pdf given by

$$f(\theta) = \begin{cases} 0 & (\theta \leq 0) \\ \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} & (0 < \theta < 1) \\ 0 & (\theta \geq 1) \end{cases}$$

where $a > 0$, $b > 0$ and $\Gamma(u)$ is the *gamma function*

$$\Gamma(u) = \int_0^\infty w^{u-1}e^{-w}\, dw = (u-1)\Gamma(u-1).$$

If $\theta \sim \text{beta}(a, b)$ then the mean of θ is

$$\text{E}(\theta) = \frac{a}{a+b}$$

and the variance of θ is

$$\text{var}(\theta) = \frac{ab}{(a+b+1)(a+b)^2}.$$

Proof: The mean is

$$\begin{aligned} \text{E}(\theta) &= \int_0^1 \theta \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\, d\theta \\ &= \frac{\Gamma(a+b)}{\Gamma(a)}\frac{\Gamma(a+1)}{\Gamma(a+b+1)}\int_0^1 \frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b)}\theta^{a+1-1}(1-\theta)^{b-1}\, d\theta \\ &= \frac{\Gamma(a+b)}{\Gamma(a)}\frac{\Gamma(a+1)}{\Gamma(a+b+1)} \\ &= \frac{a}{a+b} \end{aligned}$$

Proof continued Similarly

$$\mathrm{E}(\theta^2) = \frac{\Gamma(a+b)}{\Gamma(a)}\frac{\Gamma(a+2)}{\Gamma(a+b+2)} = \frac{(a+1)a}{(a+b+1)(a+b)}.$$

So

$$\begin{aligned}
\mathrm{var}(\theta) &= \mathrm{E}(\theta^2) - [\mathrm{E}(\theta)]^2 \\
&= \frac{(a+1)a}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2} \\
&= \frac{a(a+1)(a+b) - a^2(a+b+1)}{(a+b+1)(a+b)^2} \\
&= \frac{(a^2+a)(a+b) - a^2(a+b) - a^2}{(a+b+1)(a+b)^2} \\
&= \frac{ab}{(a+b+1)(a+b)^2}
\end{aligned}$$

8.3 Binomial observations

8.3.1 Model

Suppose that we will observe $X_1, \ldots, X_n$ where these are mutually independent given the value of θ and

$$X_i \sim \text{binomial}(N_i, \theta)$$

where $N_1, \ldots, N_n$ are known or will be observed and $\sum_{i=1}^n N_i = N$.

Then the likelihood is

$$\begin{aligned}
L(\theta;\ x) &= \prod_{i=1}^n \left(\begin{array}{c} N_i \\ x_i \end{array} \right) \theta^{x_i}(1-\theta)^{N_i - x_i} \\
&\propto \theta^S(1-\theta)^{N-S}
\end{aligned}$$

where $S = \sum_{i=1}^n x_i$.

The conjugate prior is a *beta* distribution which has a pdf proportional to

$$\theta^{a-1}(1-\theta)^{b-1}$$

for $0 < \theta < 1$.

The posterior pdf is proportional to

$$\theta^{a-1}(1-\theta)^{b-1} \times \theta^S(1-\theta)^{N-S} = \theta^{a+S-1}(1-\theta)^{b+N-S-1}.$$

This is proportional to the pdf of a beta$(a + \sum x_i,\ b + N - \sum x_i)$ distribution.

8.3.2 Example

Consider the example in section 3.4 where each animal may have a particular gene or not. The prior pdf is proportional to

$$\theta^{a-1}(1-\theta)^{b-1}$$

where we specify the values of a and b.

The posterior pdf is then proportional to

$$\theta^{a+3-1}(1-\theta)^{b+17-1}.$$

This is also the pdf of a beta distribution.

The posterior distribution does depend on the choice of prior distribution, in this case represented by the parameters a and b. To illustrate this let us consider four quite different choices of prior distribution, as follows.

Case 1 With $a = b = 1$ we obtain a uniform prior distribution for θ. In this case the posterior pdf is simply proportional to the likelihood.

Case 2 With $a = 2$ and $b = 1$ we obtain a triangular prior distribution expressing a belief that larger values of θ are more likely.

Case 3 With $a = 1$ and $b = 2$ we obtain a triangular prior distribution expressing a belief that smaller values of θ are more likely.

Case 4 With $a = b = 2$ we obtain a quadratic prior distribution expressing the belief that values of θ close to 0.5 are more likely.

Figure 18 shows, for each of these cases, the (scaled) likelihood and the prior and posterior pdf. In this way we can examine the sensitivity of the posterior distribution to the prior specification. Remember that the likelihood depends only on the data, not the prior, and is the same in all four cases. In this example, with a fairly small sample size and strongly differing prior distributions, we can see the effect of changes in the prior, although the effect is perhaps smaller than some might expect. In the more general case where we observe x animals with the gene out of n animals, the parameters of the posterior distribution are $a + x$ and $b + n - x$. Clearly, as n becomes larger, the effect of the prior parameters a and b becomes less important. If we do find that the posterior distribution is very sensitive to the choice of prior, over a reasonable range of choices, then this simply tells us that the data are not very informative and so change our beliefs relatively little.

The inferential process really consists of calculating the posterior probability distribution. Unlike non-Bayesian statistics we do not need to have "estimators", "confidence intervals" etc. However we might want to calculate particular posterior quantities for particular purposes, such as the posterior probability that the mean concentration of a compound in the bloodstream is increased after a particular treatment, or we may wish to summarise a posterior distribution simply for convenience.

For example, we might calculate the mean and variance of the posterior distribution. In this example, the *prior mean* is $a/(a+b)$ and the *prior variance* is $ab/[(a+b+1)(a+b)^2]$. The posterior mean and variance are $(a+3)/(a+b+20)$ and $(a+3)(b+17)/[(a+b+21)(a+b+20)^2]$. In the general case of observing x out of n, the posterior mean and variance are $(a+x)/(a+b+n)$ and $(a+x)(a+n-x)/[(a+b+n+1)(a+b+n)^2]$. Once again we see that the dependence on the prior will decrease as the sample size increases.

We can also calculate the probability that an unknown lies within a certain interval or range. We might have a particular interval in mind because it is of interest to us or we might simply want to do this as another way to summarise a posterior distribution. For example we might find an interval such that there is a posterior probability of 0.95 that an unknown θ lies in the interval. Such an interval is sometimes called a 95% *credible interval*. Of course there are many such intervals which we could validly calculate. Perhaps the easiest, but not necessarily the most appropriate, is a symmetric interval such that the probability that θ is less than the lower limit is equal to the probability that it is greater than the upper limit.

Table 5 gives prior and posterior means, standard deviations and symmetric 95% credible intervals for θ in the four cases above.

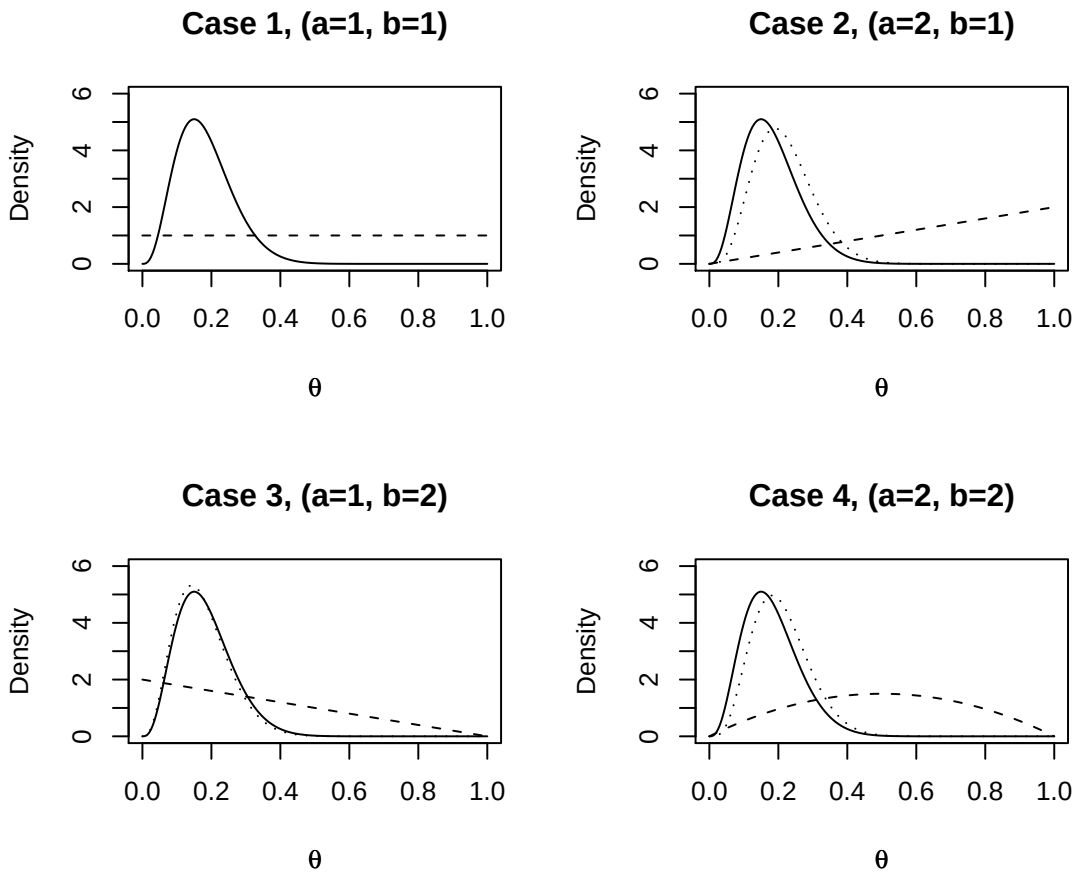

Figure 18: Likelihood (solid line), prior pdf (dashes) and posterior pdf (dots) in Cases 1, ..., 4.

			Prior			Posterior		
Case	a	b	Mean	Std.dev.	95% interval	Mean	Std.dev.	95% interval
1	1	1	0.500	0.289	$0.025 < \theta < 0.975$	0.182	0.080	$0.054 < \theta < 0.363$
2	2	1	0.667	0.236	$0.158 < \theta < 0.987$	0.217	0.084	$0.078 < \theta < 0.403$
3	1	2	0.333	0.236	$0.013 < \theta < 0.842$	0.174	0.077	$0.052 < \theta < 0.349$
4	2	2	0.500	0.224	$0.094 < \theta < 0.906$	0.208	0.081	$0.075 < \theta < 0.388$

Table 5: Numerical summaries for the four cases.

Practical 1

1: One dimensional conjugate example

We are interested in the proportion of patients given a certain drug who suffer a particular side effect. Let this proportion be θ. Our prior distribution for θ is a beta$(1, 4)$ distribution. We observe a sample of $n = 30$ patients, of whom $x = 5$ suffer the side effect. Assume that we can regard this observation as a value from the conditional binomial$(30, \theta)$ distribution of x given θ.

Use R to plot a graph showing the prior density, the likelihood and the posterior density of θ, as follows.

1. Set up a grid of θ values.

   ```
   theta<-seq(0,1,0.01)
   ```

2. Calculate the prior density.

   ```
   prior<-dbeta(theta,1,4)
   ```

3. The likelihood is proportional to $\theta^5(1-\theta)^{25} = \theta^{6-1}(1-\theta)^{26-1}$. This is proportional to the density of a beta$(6, 26)$ distribution. Calculate this.

   ```
   likelihood<-dbeta(theta,6,26)
   ```

4. The posterior distribution is beta$(1+5,\ 4+25)$. That is beta$(6, 29)$. Calculate the posterior density.

   ```
   posterior<-dbeta(theta,6,29)
   ```

5. Plot the graph.

   ```
   plot(theta,posterior,type="l",xlab=expression(theta),ylab="Density")
   lines(theta,prior,lty=2)
   lines(theta,likelihood,lty=3)
   ```

   Note that I plotted the posterior density first because I knew that this would be the tallest and that this would determine the height of the graph. Another way to achieve the same end, without having to know this, would be as follows.

   ```
   top<-max(c(prior,likelihood,posterior))
   plot(theta,prior,type="l",ylim=c(0,top),lty=2,xlab=expression(theta),
         ylab="Density")
   lines(theta,likelihood,lty=3)
   lines(theta,posterior)
   ```

2: One dimensional numerical example

This is exactly like the previous example except that we use a different prior distribution. The prior density is now proportional to

$$g(\theta) = \begin{cases} 1.1 - 2\theta & (0 < \theta < 0.5) \\ 0.1 & (0.5 \leq \theta < 1) \end{cases} .$$

1. Set up a grid of θ values.

   ```
   theta<-seq(0,1,0.01)
   ```

2. Calculate the prior density.

```
prior<-rep(0.1,101)
prior<-prior+(1-2*theta)*(theta<0.5)
```

Now, to make this a proper density we need to divide by the integral which is $0.1+(0.5\times1)/2 = 0.35$.

```
prior<-prior/0.35
```

3. The likelihood is proportional to $\theta^5(1-\theta)^{25} = \theta^{6-1}(1-\theta)^{26-1}$. This is proportional to the density of a beta$(6, 26)$ distribution. Calculate this.

```
likelihood<-dbeta(theta,6,26)
```

4. The posterior density is proportional to the prior density times the likelihood.

```
posterior<-prior*likelihood
```

We have to normalise this by dividing by the integral.

```
posterior<-posterior/(sum(posterior*0.01))
```

5. Plot the graph.

```
plot(theta,posterior,type="l",xlab=expression(theta),ylab="Density")
lines(theta,prior,lty=2)
lines(theta,likelihood,lty=3)
```

3: Two-dimensional numerical example (Weibull)

To illustrate the calculations in a two-parameter case analysed numerically we will look at inference for the parameters of a Weibull distribution using the data in Table 4.

1. The data are available in a file `weibex.txt` which can be downloaded from the module Web page.

2. To do the calculations in R we enter the data into a vector `t`.

```
t<-scan("weibex.txt")
```

(Alternatively you can simply type the numbers yourself).

3. From the module Web page, download a file called `weibpost.r` which contains a R function to evaluate the posterior density in the case where α and ρ have independent gamma prior distributions. It is written with the intention of being reasonably easy to follow rather than being the most efficient program possible. Also I have added numbers to some of the lines as comments, e.g. `# 1`, for reference.

4. Load the function into R. Type

```
weibpost<-
```

and then copy the contents of the file, paste it after this and press the "Enter" key.

5. Before we use the function we need to set up ranges of α and ρ values. For example:

```
alphgrid<-seq(0.6,1.6,0.005)
rhogrid<-seq(0.001,0.003,0.00001)
```

Do this. This will allow us to create a 201×201 grid in the α, ρ plane. The values of α go from 0.6 to 1.6 in steps of 0.005 and the values of ρ go from 0.001 to 0.003 in steps of 0.00001. Of course these numbers are more or less guesses at this stage. We may have to change them and try again.

6. We can get some idea of the sort of range to use by looking at the data. Since $F(t) = 1-\exp(-[\rho t]^\alpha)$, we find that $\ln\{-\ln[1-F(t)]\} = \alpha \ln \rho + \alpha \ln t$. So, if we could plot $\ln\{-\ln[1-F(t)]\}$ against $\ln t$, we should get a straight line with gradient α and intercept $\alpha \ln \rho$. Of course we do not know what $F(t)$ is, because we do not know the values of the parameters, but, with a moderately large sample of data like that which we have here, we can get a rough idea directly from the data. If the data arranged in increasing order are $t_{(1)}, \ldots, t_{(n)}$, then a rough estimate of $F(t_{(j)})$ is given by $(j - 0.5)/n$. We can therefore plot a graph in R as follows.

```
ltsort<-log(sort(t))
n<-length(t)
F<-((1:n)-0.5)/n
G<-log(-log(1-F))
plot(ltsort,G)
```

This gives us a line with approximate gradient 1.1 and approximate intercept -6.9. These suggest a value for α around 1.1 and a value for ρ around $\exp(-6.9/1.1) \approx 0.002$. There are methods we can use to get an idea how of wide the range around these values should be but we will leave these until later in the module.

7. Next we specify the parameters of our prior distribution. We give α a gamma$(1, 1)$ distribution and ρ independently a gamma$(3, 1000)$ distribution as follows.

```
prior<-c(1,1,3,1000)
```

8. We can now create a two-dimensional array containing values of the posterior density by typing

```
posterior<-weibpost(alphgrid,rhogrid,t,prior)
```

In the function `weibpost`, lines 1-4 are there simply to make it easier to see how the prior parameters are used. Lines 10-13 create two-dimensional arrays of α and ρ values so that every combination of values of α and ρ is represented by a position in the grid. The next line computes

$$\ln\left\{\alpha^{a_\alpha - 1} e^{-b_\alpha \alpha} \rho^{a_\rho - 1} e^{-b_\rho \rho}\right\}$$

which, apart from a constant, is the logarithm of the joint prior density. Line 18 then adds to this, for each observation,

$$\ln\left\{\alpha\rho(\rho t_i)^{\alpha-1} \exp\left[-(\rho t_i)^\alpha\right]\right\}$$

which is the contribution to the log-likelihood of observation i. Line 20 adjusts the resulting values so that they are not too large or too small before we exponentiate in line 21. Specifically we adjust so that the largest value is zero. Then, after exponentiating, the largest value is 1. This adjustment is allowed since we are going to rescale, ie normalise, the density anyway. The actual integration takes place in line 22 and the posterior density is normalised in line 23.

9. Make a contour plot of the posterior density, like figure 6, using the following command.

```
contour(alphgrid,rhogrid,posterior,xlab=expression(alpha),ylab=expression(rho))
```

We see that we have succeeded in capturing the important part of the distribution within our grid. If we had cut off an important part or if we had too much empty space we would adjust our grid and try again.

10. Produce a graph like figure 7 using the following commands.

```
alphgrid<-seq(0.6,1.6,0.01)
rhogrid<-seq(0.001,0.003,0.00002)
posterior<-weibpost(alphgrid,rhogrid,t,prior)
post<-posterior/1000
R<-rhogrid*1000
persp(alphgrid,R,post,xlab="Alpha",ylab="R",zlab="Density",phi=25,theta=45,
     ticktype="detailed",nticks=2)
```

11. To calculate the marginal posterior density of α or ρ we can integrate along the rows or columns of the array containing the posterior density. Calculate and plot the marginal density of α as follows.

```
apost<-rowSums(posterior)*rstep
plot(alphgrid,apost,type="l",xlab=expression(alpha),ylab="Density")
```

    Here `rstep` is the step size for ρ, i.e. 0.00001 in the original calculation of `posterior`.

12. Use the marginal density to calculate the posterior mean of α as follows.

```
> pmeana<-sum(alphgrid*apost)*astep
> pmeana
```

    Here `astep` is the step size for α, i.e. 0.005 in the original calculation of `posterior`.

13. Now calculate the posterior mean directly without first finding the marginal density of α, as follows.

```
> alpha<-matrix(alphgrid,nrow=101,ncol=101)
> pmeana<-sum(posterior*alpha)*0.00001*0.005
> pmeana
```

14. Each of the contours shown in figure 6 forms the boundary of a joint hpd region for α and ρ. However we have not been able to specify the probability contained within the region. Find a hpd region with a specified probability as follows. First sort the values of the posterior density into ascending order.

```
postvec<-sort(c(posterior))
```

    Next cumulatively sum them.

```
postvecsum<-cumsum(postvec)*astep*rstep
```

    Now determine the level of the posterior density which will determine the contour which we want. For example, if we want a 95% region then we want the integral over the region outside the hpd region to be 0.05.

```
crit<-max(postvec[postvecsum<0.05])
```

    Now label the points which are inside the hpd region with 1 and those outside with 0 and draw the boundary as a contour.

```
hpd<-ifelse((posterior>crit),1,0)
contour(alphgrid,rhogrid,hpd,levels=c(0.5),xlab=expression(alpha),ylab=expression(rho),
   drawlabels=FALSE)
```

4: Two-dimensional numerical example (Probit)

This refers to the example in Section 7.3.2.

1. A file `probit1.r` containing the R function shown in Figure 15 can be downloaded from the module Web page. Download this file and install the function.

   ```
   probit1<-
   ```

   Paste the contents of the file and press "Enter."

2. Enter the data.

   ```
   n<-c(560,540)
   x<-c(44,62)
   ```

3. Specify the prior.

   ```
   prior<-c(-1.24,-1.24,0.42,0.21,0.7)
   ```

4. Set up ranges for θ_1 and θ_2.

   ```
   theta1<-seq(0.01,0.99,0.01)
   theta2<-theta1
   ```

   (Because of the way the function is written, we avoid $\theta = 0$ and $\theta = 1$. We could fix this bug, of course).

5. Use the function to evaluate the posterior density.

   ```
   posterior<-probit1(theta1,theta2,n,x,prior)
   ```

6. Make a contour plot of the posterior density.

   ```
   contour(theta1,theta2,posterior$density)
   ```

   (The result of the function is a list. One of the elements is the posterior density).

7. We see that the posterior probability is concentrated in one region so we adjust the ranges for θ_1 and θ_2 and try again.

   ```
   theta1<-seq(0.05,0.15,0.001)
   theta2<-seq(0.05,0.15,0.001)
   post<-probit1(theta1,theta2,n,x,prior)
   dens<-post$density
   contour(theta1,theta2,dens,xlab=expression(theta[1]),ylab=expression(theta[2]))
   ```

8. This time the result looks more satisfactory. We can add a line showing where $\theta_1 = \theta_2$.

   ```
   abline(0,1,lty=2)
   ```

   We can easily see that nearly all of the posterior probability lies in the region where $\theta_2 > \theta_1$.

9. The other element of the result of the function is the posterior probability that $\theta_2 > \theta_1$. Extract this.

   ```
   posterior$prob
   ```

10. To make the plot of the prior and posterior densities of the log relative risk, downsload the function `probit2` in the file `probit2.r` from the module Web page and install it in R.

```
probit2<-
```

Paste the contents of the file.

11. Set up the data and prior. (They are the same as before so you should already have these).

```
n<-c(560,540)
x<-c(44,62)
prior<-c(-1.24,-1.24,0.42,0.21,0.7)
```

12. Set up ranges for γ, the log relative risk, and θ_2.

```
gamma<-seq(-1,1,0.02)
theta2<-seq(0.05,0.15,0.001)
```

13. Use the function.

```
probit2(gamma,theta2,n,x,prior)
```

Reference

Anturane Reinfarction Trial Research Group, 1980. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. *New England Journal of Medicine*, **302**, 250-256.