MAS3301 Bayesian Statistics

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 2, 2008-9

5 A Model with Two Unknowns

5.1 Deterministic and stochastic models

We can apply the ideas of probability to "mathematical" models. We often distinguish between "deterministic" models, in which random variables do not play a part in the model itself, and "stochastic" models, in which they do. Clearly probability is involved in stochastic models but it is also relevant in deterministic models when there is uncertainty about parameter values and when observations are made with error.

5.2 More than one unknown: Basic ideas

So far we have mostly looked at Bayesian inference in the case where we have a single unknown quantity, usually a parameter. Now we will look at what happens when we have more than one unknown parameter. The principle is the same when we have more than one parameter. We simply obtain a joint posterior distribution for the parameters. For example, if there are two parameters, we might produce a contour plot of the posterior pdf, as shown in figure 6, or a "3-d" plot, as shown in figure 7. If there are more than two parameters we need to "integrate out" some of the parameters in order to produce graphs like this.

As usual, the basic rule is **posterior** $\propto$ **prior** $\times$ **likelihood.** If necessary, the normalising constant is found by integrating over all parameters. Posterior means, variances, marginal probability density functions, predictive distributions etc. can all be found by suitable integrations. In practice the integrations are often carried out numerically by computer. Apart from being the only practical means in many cases, this removes the pressure to use a convenient conjugate prior.

Sometimes our beliefs might be represented by a model containing several parameters and we might want to answer questions about a number of them. For example, in a medical experiment, we might be interested in the effect of a new treatment on several different outcome measures so we might want to make inferences about the change in the mean for each of these when we move from the old to the new treatment. In frequentist statistics this can give rise to the "multiple testing problem." This problem does not arise for Bayesians. For a Bayesian the inference always consists of the posterior distribution. Once we have calculated the posterior distribution we can calculate whatever summaries we want from it without any logical complications. For example, we could calculate a posterior probability that the mean outcome measure has increased from one treatment to the other for each outcome, or a joint probability that it has increased for every member of some subset of the outcomes or any or all of many other summaries.

5.3 The bivariate normal distribution

The normal distribution can be extended to deal with two variables. (In fact, we can extend this to more than two variables).

If Y_1 and Y_2 are two continuous random variables with joint pdf

$$f(\underline{y}) = (2\pi)^{-1}|V|^{-1/2}\exp\left\{-\frac{1}{2}(\underline{y}-\underline{\mu})^T V^{-1}(\underline{y}-\underline{\mu})\right\}$$

for $-\infty < y_1 < \infty$ and $-\infty < y_2 < \infty$ then we say that Y_1 and Y_2 have a bivariate normal distribution with *mean vector* $\underline{\mu} = (\mu_1,\ \mu_2)^T$ and *variance matrix*

$$V = \begin{pmatrix} v_{1,1} & v_{1,2} \\ v_{1,2} & v_{2,2} \end{pmatrix}$$

where μ_1 and μ_2 are the means of Y_1 and Y_2 respectively, $v_{1,1}$ and $v_{2,2}$ are their variances, $v_{1,2}$ is their covariance and $|V|$ is the determinant of V.

If Y_1 and Y_2 are independent then $v_{1,2} = 0$ and, in the case of the bivariate normal distribution, the converse *is* true.

Note that, if X and Y both have normal marginal distributions it does not necessarily follow that their joint distribution is bivariate normal, although, in practice, the joint distribution often is bivariate normal. However, if X and Y both have normal distributions and are independent then their joint distribution is bivariate normal with zero covariance.

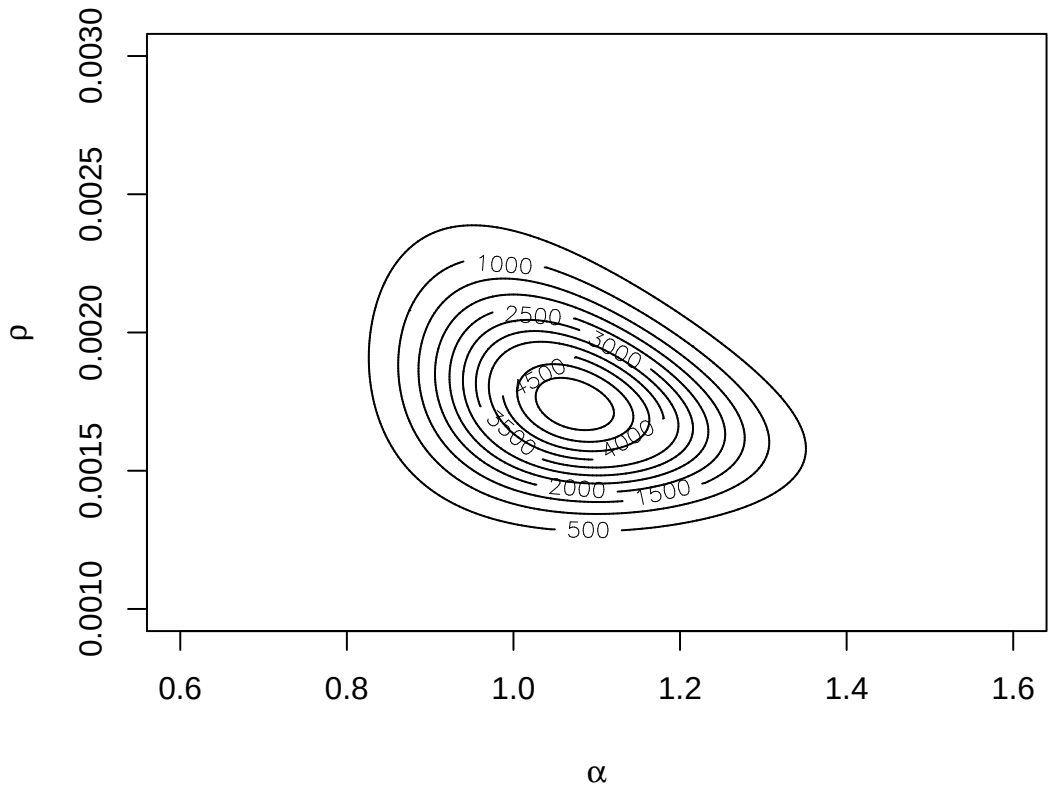

Figure 6: Posterior density of two unknowns: Contour plot

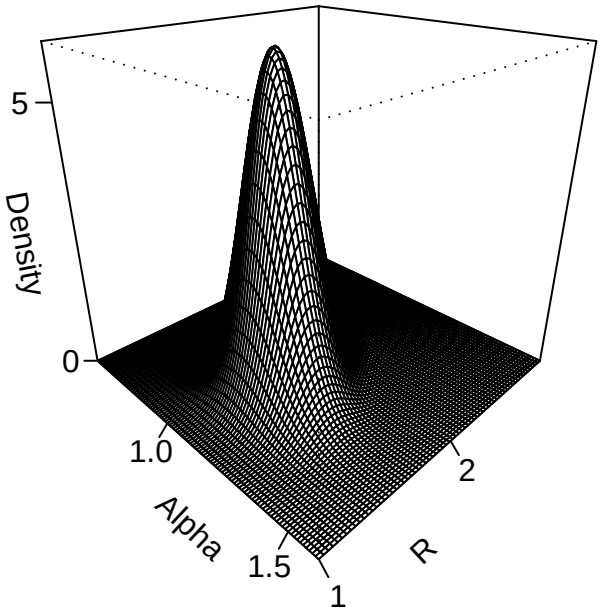

Figure 7: Posterior density of two unknowns: Wireframe plot

If Y_1 and Y_2 have a bivariate normal distribution then $a_1Y_1 + a_2Y_2$ is also normally distributed, where a_1 and a_2 are constants. For example, if $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ and X and Y are independent then $X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

5.4 Functions of continuous random variables (Revision)

5.4.1 Theory

As we shall see in the example below, we sometimes need to find the distribution of a random variable which is a function of another random variable. Suppose we have two random variables X and Y where $Y = g(X)$ for some function $g()$. In this section we will only consider the case where $g()$ is a strictly monotonic, i.e. either strictly increasing or strictly decreasing, function.

Suppose first that $g()$ is a strictly increasing function so that if $x_2 > x_1$ then $y_2 = g(x_2) > y_1 = g(x_1)$. In this case the distribution functions $F_X(x)$ and $F_Y(y)$ are related by

$$F_Y(y) = \Pr(Y < y) = \Pr(X < x) = F_X(x).$$

We can find the relationship between the probability density functions , $f_Y(y)$ and $f_X(x)$, by differentiating with respect to y. So

$$f_Y(y) = \frac{d}{dy}F_y(y) = \frac{d}{dy}F_X(x) = \frac{d}{dx}F_X(x) \times \frac{dx}{dy} = f_X(x)\frac{dx}{dy} = f_X(x)\left(\frac{dy}{dx}\right)^{-1}.$$

Similarly, if $g()$ is a strictly decreasing function so that if $x_2 > x_1$ then $y_2 = g(x_2) < y_1 = g(x_1)$,

$$F_Y(y) = \Pr(Y < y) = \Pr(X > x) = 1 - F_X(x)$$

and

$$f_Y(y) = -f_X(x)\frac{dx}{dy}$$

but here, of course, dx/dy is negative.

So, if $g()$ is a strictly monotonic function

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| \text{ where } \left|\frac{dx}{dy}\right| \text{ is the modulus of } \frac{dx}{dy}.$$

A simple way to remember this is to remember that an element of probability $f_X(x)\delta x$ is preserved through the transformation so that (for a strictly increasing function)

$$f_Y(y)\delta y = f_X(x)\delta x.$$

5.4.2 Example

Suppose for example that $X \sim N(\mu,\ \sigma^2)$ and that $Y = e^X$. So $X = \ln(Y)$ and $dx/dy = y^{-1}$. Now

$$f_X(x) = (2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \qquad (-\infty < x < \infty)$$

so

$$f_Y(y) = \frac{1}{y}(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{\ln(y)-\mu}{\sigma}\right)^2\right\} \qquad (0 < y < \infty).$$

The resulting distribution for Y is called a *lognormal* distribution because $\ln(Y)$ has a normal distribution. It can be useful for representing beliefs about quantities which can only take positive values.

5.5 Deterministic model example

As an example, suppose that, after material is extracted from an organism, the concentration of a certain compound decreases exponentially over time so that the concentration Z, in mmol/l, after t minutes is given by

$$Z = Z_0 e^{-ct} = e^{a-ct},$$

where $Z_0 = e^a$ is the initial concentration and c is a constant decay rate.

There are two parameters, a and c. We might describe our beliefs about a and c by a bivariate normal distribution where the mean and variance of a are μ_a and σ_a^2, the mean and variance of c are μ_c and σ_c^2 and the covariance of a and c is γ. Then $a - ct$ is normal with mean

$$\mathrm{E}(a - ct) = \mu_a - t\mu_c$$

and variance

$$\mathrm{var}(a - ct) = \sigma_a^2 + t^2\sigma_c^2 - 2t\gamma.$$

Thus we can find probabilities for values of Z at time t.

Suppose our beliefs about a and c are such that a and c are independent so $\gamma = 0$. Suppose our median for Z_0 is 100 so $\mu_a = \ln(100) = 4.605$ and we are 95% sure that $50 < Z_0 < 200$ so $\sigma_a = \ln(2)/1.96 = 0.3536$ and $\sigma_a^2 = 0.1251$. The figure 1.96 arises because the probability that a

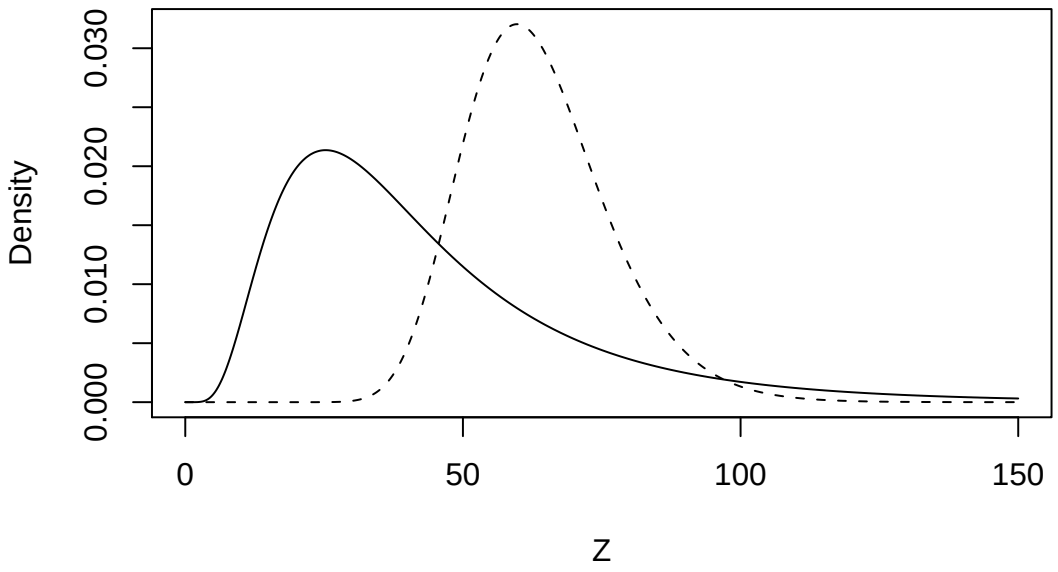

Figure 8: P.d.f. for Z at $t = 100$. Solid line: original. Dashed line: after observation at $t = 50$.

normal random variable is within ± 1.96 standard deviations of its mean is 0.95. Similarly suppose we choose $\mu_c = 0.01, \sigma_c^2 = 0.000025$ (so we are 95% sure that $0.0002 < c < 0.0198$). Then

$$\ln(Z) \sim N(4.605 - 0.01t, 0.1251 + 0.000025t^2).$$

For example, at $t = 100$, we have $\ln(Z) \sim N(3.605, 0.3751)$ and we are 95% sure that $2.405 < \ln(Z) < 4.805$, that is $11.07 < Z < 122.17$.

The density for Z at $t = 100$ is shown by the solid line in figure 8.

We may prefer to give a hpd interval. The earlier 95% interval was a hpd interval for $\ln(Z)$. It is more tricky to find a hpd interval for Z in this example. An exact solution for a given probability requires numerical iteration. However we can easily find the probability for a given interval. The density is approximately equal at the points $Z = 10$ and $Z = 65$. Now

$$\begin{aligned}
\Pr(10 < Z < 65) &= \Pr(2.3026 < \ln(Z) < 4.1744) \\
&= \Pr\left(\frac{2.3026 - 3.605}{\sqrt{0.3751}} < W < \frac{4.1744 - 3.605}{\sqrt{0.3751}}\right) \\
&= \Pr(-2.127 < W < 0.9297) \\
&\simeq 0.807
\end{aligned}$$

or 80.7% (where $W \sim N(0, 1)$).

We know that a 95% hpd interval has two properties which we can use to find the interval. Let the lower and upper limits be l_1 and l_2 and the pdf in question be $f(x)$. Then, for a 95% hpd interval:

$$\int_{l_1}^{l_2} f(x)\,dx = 0.95 \quad \text{and} \quad f(l_1) = f(l_2).$$

We can use R to find a h.p.d. interval for Z. The distribution of Z is actually a lognormal distribution. In R the parameters to specify this are the mean and standard deviation of $\log(Z)$. In this case these are 3.605 and $\sqrt{0.3751} = 0.6125$ respectively. We can write a function which, if we guess a lower limit for an interval, will tell us the upper limit and the density values at the two limits. Here is such a function definition.

```
hpd<-function(lower,mean,var,prob)
{stddev<-sqrt(var)
 plower<-plnorm(lower,mean,stddev)
 pupper<-plower+prob
 upper<-qlnorm(pupper,mean,stddev)
 ldens<-dlnorm(lower,mean,stddev)
 udens<-dlnorm(upper,mean,stddev)
 result<-data.frame(lower,upper,ldens,udens)
 result
}
```

The R function `plnorm` evaluates the distribution function, `dlnorm` evaluates the pdf and `qlnorm` evaluates a quantile. That is it finds a value q such that $F(q)$ takes a specified value, where $F()$ is the distribution function.

Here is an example of our function's use.

```
> m<-3.605
> v<-0.3751
> p<-0.95
> hpd(9,m,v,p)
  lower    upper       ldens       udens
1     9 108.0604 0.005155869 0.001281801
> hpd(5,m,v,p)
  lower    upper       ldens       udens
1     5 101.0626 0.000644929 0.001651351
> hpd(7,m,v,p)
  lower    upper       ldens       udens
1     7 102.8224 0.002372777 0.001548699
> hpd(6,m,v,p)
  lower    upper       ldens       udens
1     6 101.6593 0.001356244 0.001615746
> hpd(6.5,m,v,p)
  lower    upper       ldens      udens
1   6.5 102.1545 0.001827710 0.00158683
> hpd(6.3,m,v,p)
  lower    upper       ldens       udens
1   6.3 101.9375 0.001630046 0.001599429
> hpd(6.2,m,v,p)
  lower    upper       ldens       udens
1   6.2 101.8388 0.001535725 0.001605202
> hpd(6.25,m,v,p)
  lower    upper       ldens       udens
1  6.25 101.8874 0.001582506 0.001602358
```

We see that $6.25 < Z < 101.89$ is approximately a 95% h.p.d. interval for Z. If we wished, we could write a more developed function which did not require us to use trial and error.

Clearly, if we observe Z at two time points we can determine the values of both parameters. If we only make one observation this reduces, but does not eliminate, our uncertainty. For example, suppose in our example that, at $t = 50$, we observe $Z = 90$. This means that $a - 50c = \ln(90) = 4.500$. It can be shown that, conditional on this information, a and c have a bivariate normal distribution in which the mean of a is $\tilde{\mu}_a = 4.8684$, the mean of c is $\tilde{\mu}_c = 0.007368$, the variance of a is $\tilde{\sigma}_a^2 = 4.168 \times 10^{-2}$, the variance of c is $\tilde{\sigma}_c^2 = 1.667 \times 10^{-5}$ and the covariance of a and c is $\tilde{\gamma} = 8.336 \times 10^{-4}$. The distribution for $\ln(Z)$ at $t = 100$ is now normal with mean 4.1316 and variance 0.04168. The density of the new distribution for Z at $t = 100$ is shown by the dashed line in figure 8. (We omit, for now, the details of how to find the new distribution after an observation).

5.6 Deterministic Example with Error

Suppose in our deterministic example above that we can not observe the actual value of Z at time t but instead observe

$$\tilde{Z}_t = \exp\{a - ct + \varepsilon_t\},$$

where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ and the value of σ_ε^2 is known. (This would not usually be the case but we assume here that we know the properties of our measuring instrument sufficiently well). Thus our observation contains multiplicative errors. Assume further that the errors at different times are independent. Suppose we make observations $\tilde{Z}_1, \ldots, \tilde{Z}_n$ on the population at times $t_1, \ldots, t_n$ and let $Y_i = \ln \tilde{Z}_i$. If our *prior* probability distribution for a, c, which represents our beliefs before we see the data, is bivariate normal as before then *a priori,* our joint distribution for $a, c, Y_1, \ldots, Y_n$ is *multivariate normal*.

We specify a *multivariate* normal distribution, that is a normal distribution for several variables, by a mean vector $\underline{\mu}$ and a variance matrix V. This is a straightforward generalisation of the *bivariate* normal distribution seen in section 5.3. Let

$$\underline{\beta} = \begin{pmatrix} a \\ c \end{pmatrix}.$$

In this way, *a priori,* $\underline{\beta} \sim N(\underline{\mu}_0, V_0)$, where

$$\underline{\mu}_0 = \begin{pmatrix} \mu_a \\ \mu_c \end{pmatrix}$$

and

$$V_0 = \begin{pmatrix} \sigma_a^2 & \gamma \\ \gamma & \sigma_c^2 \end{pmatrix}.$$

It can be shown that our conditional distribution for a, c given $Y_1 = y_1, \ldots, Y_n = y_n$ is bivariate normal with parameters given by the following formulae. (Proof omitted. Consider these as "given" for this example). This new distribution, after we have seen the data, is called our *posterior* distribution.

Let X be a $n \times 2$ matrix, the first column of which is $(1, 1, 1, \ldots, 1)^T$ and the second column of which is $(-t_1, -t_2, \ldots, -t_n)^T$. Let $\underline{y} = (y_1, y_2, \ldots, y_n)^T$.

Let

$$\underline{\hat{\beta}} = (X^T X)^{-1} X^T \underline{y} = (\hat{a},\ \hat{c})^T,$$

$$\text{where} \qquad \hat{c} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{a} = \bar{y} - \hat{c}\bar{x}, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \qquad \text{and} \qquad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Some readers will recognise this as the least squares estimate of $\underline{\beta}$. That is it is the "usual" estimates of a (intercept) and c (gradient) in a simple regression model.

Then the posterior distribution of $\underline{\beta}$ is bivariate normal with mean $\underline{\mu}_1$ and variance V_1 where

$$
\begin{aligned}
V_1 &= (V_0^{-1} + \sigma^{-2} X^T X)^{-1} \\
\underline{\mu}_1 &= V_1(V_0^{-1}\underline{\mu}_0 + \sigma^{-2} X^T X \underline{\hat{\beta}}) = V_1(V_0^{-1}\underline{\mu}_0 + \sigma^{-2} X^T \underline{y})
\end{aligned}
$$

Notice that the *posterior mean*, the mean of the posterior distribution, is a weighted average of the prior mean and the least squares estimate in this case.

Suppose that the data are as follows and $\sigma_\varepsilon^2 = 0.0025$.

Time t	25	50	75	100	125	150
Measured Concentration $\tilde{Z}$	113	81	74	52	43	36
Log Concentration Y	4.73	4.39	4.30	3.95	3.76	3.58

The posterior means of a and c are now 4.913 and 9.091×10^{-3} respectively and their posterior variances are 2.114×10^{-3} and 2.234×10^{-7}. The posterior covariance of a and c is 1.948×10^{-5}.

6 Simple Numerical Methods

6.1 The need for integration

Consider a general Bayesian inference problem. We have a (scalar or vector) unknown θ. Usually θ will be a "parameter," or a vector of parameters. We have observations Y. Let us assume that both θ and Y are continuous so that we can talk in terms of densities and integrals. If one is discrete then we replace the probability density with a probability. If θ is discrete then we replace the integrals with summations.

We have a prior probability density function $f_\theta^{(0)}(\theta)$ for θ. Let the range of possible values of θ be Θ.

We have a sampling distribution pdf $f_{Y|\theta}(y \mid \theta)$ for the observations Y. (We are treating Y as a vector here so this is the joint pdf for all of the elements of Y).

The joint density of θ and Y is then $f_\theta^{(0)}(\theta) f_{Y|\theta}(y \mid \theta)$.

The posterior density of θ is

$$
f_\theta^{(1)}(\theta) = \frac{f_\theta^{(0)}(\theta) f_{Y|\theta}(y \mid \theta)}{\int_\Theta f_\theta^{(0)}(\theta) f_{Y|\theta}(y \mid \theta)\ d\theta}.
$$

Let $g(\theta)$ be some scalar function of θ. Then the posterior mean of $g(\theta)$ is

$$
\mathrm{E}^{(1)}[g(\theta)] = \frac{\int_\Theta g(\theta) f_\theta^{(0)}(\theta) f_{Y|\theta}(y \mid \theta)\ d\theta}{\int_\Theta f_\theta^{(0)}(\theta) f_{Y|\theta}(y \mid \theta)\ d\theta}.
$$

To find the posterior variance of $g(\theta)$ we can also find $\mathrm{E}^{(1)}[\{g(\theta)\}^2]$ and then $\mathrm{var}^{(1)}[g(\theta)] = \mathrm{E}^{(1)}[\{g(\theta)\}^2] - \{\mathrm{E}^{(1)}[g(\theta)]\}^2$.

Each integral which we need here is of the form

$$
\int_\Theta g(\theta) f_\theta^{(0)}(\theta) f_{Y|\theta}(y \mid \theta)\ d\theta
$$

(where we might have $g(\theta) = 1$).

Sometimes, when we have conjugate priors, these integrals work out analytically. Often they do not so we use numerical methods.

Note that θ is often a vector so we have multiple integrations. In this section we will look just at the scalar case. We will consider the use of multiple integration in lecture 7.

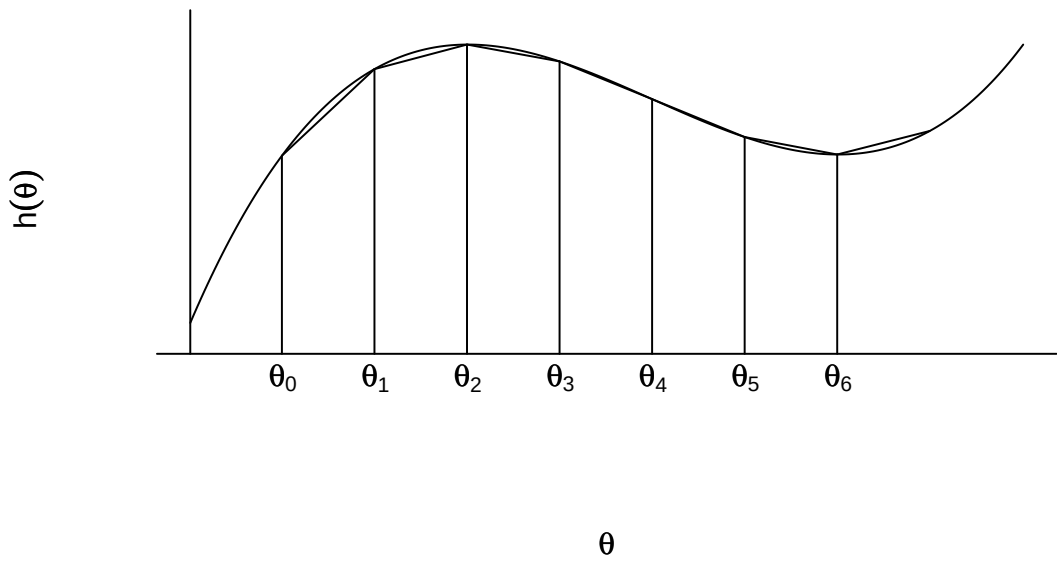

Figure 9: Numerical integration using the trapezium rule.

6.2 Trapezium rule

Various more refined methods are available but a simple trapezium rule is often sufficient.

Consider a 1-dimensional case (i.e. θ a scalar).

A simple version of the procedure is as follows.

1. Choose lower and upper limits, $a,\ b$, for the integration. In some cases we might want an integral where one of the limits is infinite, or perhaps both are. In such cases we need to choose suitable finite limits in such a way that the function beyond the limits would contribute little to the integral.

2. Set up an array of values $\theta_0, \theta_1, \ldots, \theta_m$ of values for θ. These values are separated by a *step size* of $\delta\theta$. We assume (for now) that the steps are all equal. That is $\theta_{j+1} = \theta_j + \delta\theta$. We set $\theta_0 = a$ and $\theta_m = b$ so $\delta\theta = (b-a)/m$.

3. Evaluate $h(\theta) = g(\theta) f_\theta^{(0)}(\theta) f_{Y|\theta}(y \mid \theta)$ at each value θ_j.

4. Approximate the area under the curve by the total area of a set of m trapezium-shaped columns. Column j stands on the θ axis and has as its base the interval $(\theta_{j-1},\ \theta_j)$. It has vertical sides with heights $h(\theta_{j-1})$ and $h(\theta_j)$. The top of the column is a straight line joining the tops of the sides. Each column has width $\delta\theta$. The area of column j is $\delta\theta \times [h(\theta_{j-1}) + h(\theta_j)]/2$.

$$\int_\Theta h(\theta)\ d\theta \approx \sum_{j=1}^{m} \frac{h(\theta_{j-1}) + h(\theta_j)}{2}\delta\theta = \sum_{j=0}^{m} h(\theta_j)\ \delta\theta - \frac{h(\theta_0) + h(\theta_m)}{2}\delta\theta \tag{2}$$

See figure 9.

Usually the density becomes close to zero in the tails of the distribution so edge effects are often not a problem. In such cases we can ignore the final term $\{[h(\theta_0) + h(\theta_m)]/2\}\delta\theta$ in 2 and use

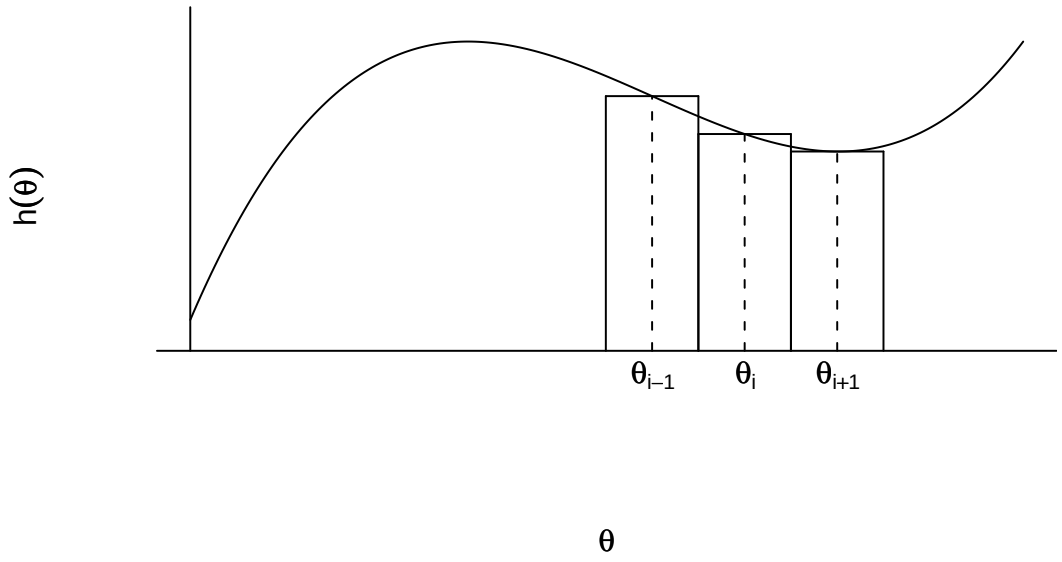

Figure 10: Numerical integration using rectangular columns.

$$\int_{\Theta} h(\theta)\ d\theta \approx \sum_{j=0}^{m} h(\theta_j)\ \delta\theta. \tag{3}$$

We can interpret this as approximating the integral by the total area of a set of rectangular columns where column j has height $h(\theta_j)$ and its base is the interval $(\theta_j - \delta\theta/2,\ \theta_j + \delta\theta/2)$. See figure 10.

Note:

1. We often need some trial and error to find an appropriate range for the grid. Plotting $h(\theta)$ against θ can help.

2. Of course we need to choose a value for m and for the step size $\delta\theta$. Smaller step sizes will tend to give more accurate answers but increasing the number of evaluations increases the computational time. Once we have settled on the lower and upper limits between which we are going to integrate numerically, we can try increasing m, and therefore decreasing $\delta\theta$, to see whether this changes the answer more than a negligible amount. If necessary, we can do our final evaluation using a large value of m even if this requires waiting a few minutes for the computer to complete the calculation.

3. Care needs to be taken with numerical issues. Values of functions can often be very small. It is usually best to evaluate logs first. See the examples below.

6.3 Example: Chester Road

Consider the Chester Road data described in section 4.4. To illustrate the principles involved when we use a non-conjugate prior, we will try using a "triangular" prior distribution with the pdf

$$f^{(0)}(\lambda) = \begin{cases} 4(1-2\lambda) & (0 < \lambda < 0.5) \\ 0 & (\text{otherwise}) \end{cases}$$

The calculations are actually quite simple in this case but we will do them using a general method. The likelihood is proportional to

$$\lambda^{115} e^{-1200\lambda}.$$

It is usually better to work in terms of logarithms, to avoid very large and very small numbers. The log likelihood, apart from an additive constant, is

$$115 \ln(\lambda) - 1200\lambda.$$

We will also take logarithms of the prior density. Note that we could not do this where the prior density is zero but we really do not need to. In this example the logarithm of the prior density is, apart from an additive constant,

$$\ln(1 - 2\lambda).$$

Having added the log prior to the log likelihood we subtract the maximum of this so that the maximum becomes zero. We then take exponentials and the maximum of this will be 1.0. Again, the reason for doing this is to avoid very large or very small numbers. We then normalise by finding the integral and dividing by this. The integral is found numerically using a simple trapezium rule.

Suitable R commands are as follows. The likelihood is, of course, the same as in section 4.4 and this allows us to choose sensible limits for the numerical integration, in this case 0.05 and 0.15. There is no point in evaluating the integrand at lots of points where it is close to zero. Sometimes we might need to use a bit of trial and error until we find a suitable range, plotting a graph of the results each time to see whether we have either cut off an important part of the function at either end or included too much empty space.

```
stepsize<-0.001
lambda<-seq(0.05,0.15,stepsize)
prior<-4*(1-2*lambda)
logprior<-log(1-2*lambda)
loglik<-115*log(lambda)-1200*lambda
logpos<-logprior+loglik
logpos<-logpos-max(logpos)
posterior<-exp(logpos)
posterior<-posterior/(sum(posterior)*stepsize)
plot(lambda,posterior,type="l",ylab="Density")
lines(lambda,prior)
```

We can, of course, get R to calculate the statistics we need for the likelihood from the raw data. In this case it is actually just the number of vehicles which passed.

We can compare our results with those obtained in section 4.4. Figure 11 shows the prior and posterior densities and the (scaled) likelihood and can be compared to figure 5. We see that the posterior density is now even closer to the likelihood and is, in fact, almost indistinguishable from it. The R commands which I used to produce this plot are as follows.

```
plot(lambda,posterior,type="l",xlab=expression(lambda),ylab="Density")
like<-dgamma(lambda,116,1200)
lines(lambda,like,lty=2)
lines(lambda,prior,lty=3)
abline(h=0)
```

In this example, of course, the likelihood is proportional to a gamma(116, 1200) density.

Having found the posterior density of λ we can easily find, for example, the posterior mean, variance and standard deviation, using R as follows.

```
> postmean<-sum(posterior*lambda)*stepsize
> postmean
[1] 0.09646694
> postvar<-sum(posterior*(lambda^2))*stepsize-postmean^2
> postvar
[1] 8.01825e-05
> sqrt(postvar)
[1] 0.008954468
```

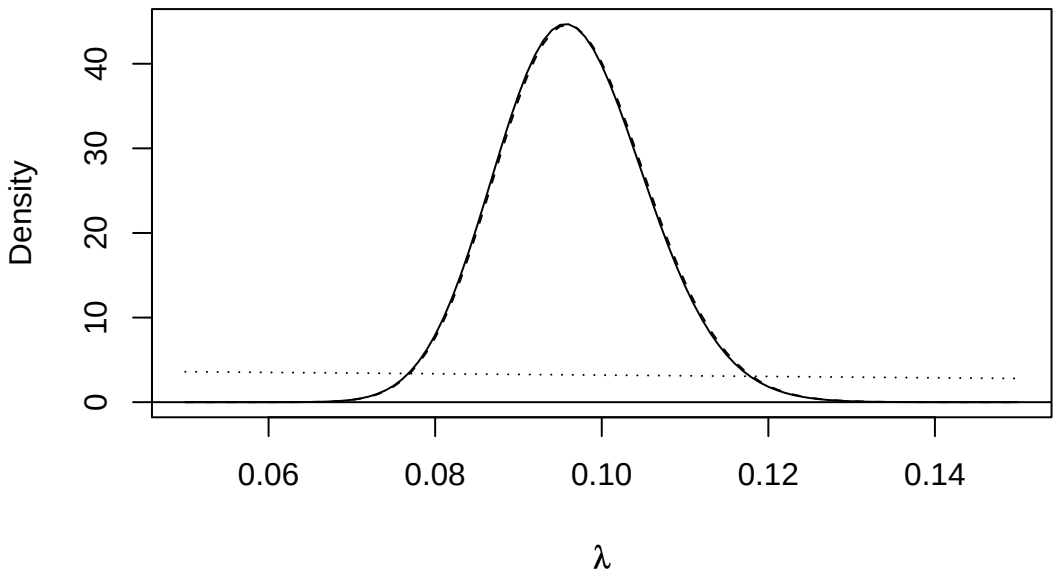

Figure 11: Chester road traffic arrival rate, "triangular" prior. Dots: prior pdf, Dashes: scaled likelihood, Solid line: posterior pdf.

6.4 Example: The upper limit for a continuous uniform distribution

Suppose that we will make observations $X_1, \ldots, X_n$ where, given the value of the parameter θ, these are independent and each has a continuous uniform distribution on the interval $(0, \theta)$ but the value of θ is unknown. Our prior distribution for θ is a gamma(4,0.5) distribution. Our prior mean is thus $4/0.5 = 8$ and our prior variance is $4/0.5^2 = 16$, giving a standard deviation of 4.

Our data, with $n = 15$, are as follows.

```
5.21 2.76 2.22 2.36 3.22 2.72 5.34 5.33 1.93 0.99 0.54 1.47 3.06 1.51 3.87
```

The largest observation is $x_{\max} = 5.34$. Clearly the likelihood is zero for $\theta < x_{\max}$ since the probability of observing $x_{\max} > \theta$ is zero. For $\theta > x_{\max}$ the likelihood is

$$L(\theta) = \prod_{i=1}^{15} \frac{1}{\theta} = \theta^{-15}.$$

The posterior density is therefore proportional to

$$g(\theta) = \begin{cases} 0 & (\theta \leq 5.34) \\ \theta^3 e^{-\theta/2} \times \theta^{-15} = \theta^{-12} e^{-\theta/2} & (\theta > 5.34) \end{cases}$$

For values of $\theta > 5.34$, the log posterior density is, apart from an additive constant,

$$-12 \ln(\theta) - \theta/2.$$

We can find the posterior density using the following R commands. The lower limit for the integration is, naturally, 5.34 and we make the allowance for the boundary here. The upper limit has been chosen to be 8.34. The initial plot shows that this was a reasonable choice as $g(\theta)$ has become very small at this point.

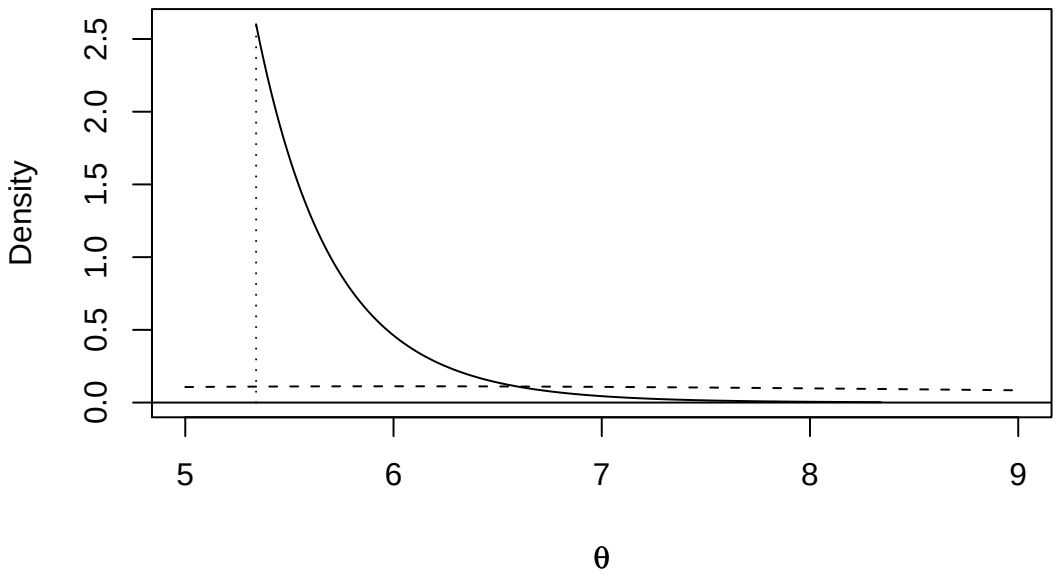

Figure 12: Posterior density (solid line) and prior density (dashes) for the upper limit of a continuous uniform distribution.

```
> theta<-seq(5.34,8.34,0.01)
> lg<--12*log(theta)-theta/2
> lg<-lg-max(lg)
> g<-exp(lg)
> plot(theta,g,type="l")
> posterior<-g/((sum(g)-g[1]/2)*0.01)
```

Figure 12 shows the posterior density and the prior density.

We can calculate the posterior mean, variance and standard deviation as follows.

```
> postmean<-(sum(posterior*theta)-posterior[1]*theta[1]/2)*0.01
> postmean
[1] 5.74286
> postvar<-(sum(posterior*(theta^2))-posterior[1]*(theta[1]^2)/2)*0.01-postmean^2
> postvar
[1] 0.1711295
> sqrt(postvar)
[1] 0.4136781
```

6.5 Example: Acceptance sampling

A batch of m items is manufactured. Each of these items might be defective. A simple random sample of $n < m$ of the items in the batch is selected and inspected. We find r defectives in the sample. What does this tell us about the number of defectives d in the batch?

First we need a prior distribution for d. Notice that d has to be an integer so we need a discrete prior distribution. There are, of course, many possibilities but one very simple possibility is a *discrete uniform distribution* on the interval $[0, m]$. This gives an equal probability to each possible number of defectives in the batch. We will consider a different, perhaps more realistic, prior as an exercise at the end of this chapter.

Let a and b be integers such that $a < b$. If a random variable X is equally likely to take any integer value between a and b inclusive and can not take any other value then we say that X has a discrete uniform distribution on $[a, b]$.

The probabilities are as follows.

$$\Pr(X = i) = \begin{cases} 0 & i < a \\ 1/(b-a+1) & a \leq i \leq b \\ 0 & i > b \end{cases}$$

where i is an integer.

It is easy to show that the mean of X is $(a+b)/2$ and the variance is $(b-a)(b-a+2)/12$.

In our case $a = 0$ and $b = m$ so the probabilities are all $(m+1)^{-1}$, the mean is $m/2$ and the variance is $m(m+2)/12$.

The likelihood is obviously zero for $d < r$. For $r \leq d \leq m$, the likelihood is

$$L(d) = \frac{\binom{d}{r}\binom{m-d}{n-r}}{\binom{m}{n}}. \tag{4}$$

Given d, the number of defectives in the sample has a *hypergeometric distribution* and this is a probability from that distribution. We can explain (4) as follows. We choose a random sample of n items from the batch of m. Since it is a simple random sample, all possible samples are equally likely. The denominator on the right of (4) is the number of possible different samples which we could choose. The numerator is the number of possible samples which contain exactly r defectives. It is the product of the number of ways to choose the r defectives in the sample out of the total of d available defectives and the number of ways to choose the $n - r$ non-defectives in the sample out of the total of $m - d$ available non-defectives.

Since the prior probabilities are equal for all values of d, the posterior probabilities are proportional to the likelihood. Fortunately for us, R has a function to calculate hypergeometric probabilities and all we have to do to find the posterior probabilities is normalise the hypergeometric probabilities by dividing by their total.

Suppose, for example, that $m = 100,\ n = 20$ and $r = 2$. The following R commands will do the necessary calculations.

```
d<-2:100
e<-100-d
g<-dhyper(2,d,e,20)
post<-g/sum(g)
```

Figure 13 shows the posterior probabilities. It seems that there may be quite a large proportion of defectives in this batch! We might like to find the posterior probability that there are more than 20 defectives in the batch.

```
> sum(post[d>20])
[1] 0.1271451
```

We see that we have a posterior probability of 0.127 for this event.

Suppose that we observed no defectives in our sample. This, of course, makes smaller numbers of defectives in the batch more likely. How likely is it that there are no defectives in the batch?

```
> d<-0:100
> e<-100-d
> g<-dhyper(0,d,e,20)
> post<-g/sum(g)
> post[1]
[1] 0.2079208
```

We see that, given no defectives in our sample, we would have a posterior probability of 0.208 for the event that there are no defectives in the batch.

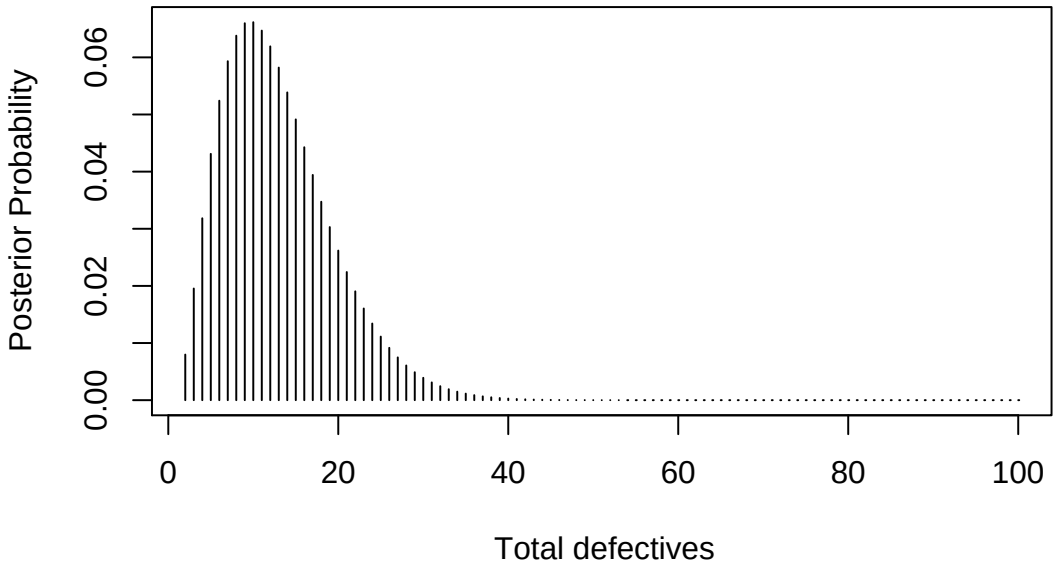

Figure 13: Posterior probabilities for numbers of defectives in a batch of 100 from a sample of 20 containing 2 defectives.

6.6 Problems 2

Useful integrals: In solving these problems you might find the following useful.

- Gamma functions: Let a and b be positive. Then

$$\int_0^\infty x^{a-1}e^{-bx}\,dx = \frac{\Gamma(a)}{b^a}$$

  where

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}\,dx = (a-1)\Gamma(a-1).$$

  If a is a positive integer then $\Gamma(a) = (a-1)!$.

- Beta functions: Let a and b be positive. Then

$$\int_0^1 x^{a-1}(1-x)^{b-1}\,dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

1. We are interested in the mean, λ, of a Poisson distribution. We have a prior distribution for λ with density

$$f^{(0)}(\lambda) = \begin{cases} 0 & (\lambda \le 0) \\ k_0(1+\lambda)e^{-\lambda} & (\lambda > 0) \end{cases}.$$

   (a) i. Find the value of k_0.
   ii. Find the prior mean of λ.
   iii. Find the prior standard deviation of λ.

   (b) We observe data $x_1, \ldots, x_n$ where, given λ, these are independent observations from the Poisson(λ) distribution.

i. Find the likelihood.
ii. Find the posterior density of λ.
iii. Find the posterior mean of λ.

2. We are interested in the parameter, θ, of a binomial(n, θ) distribution. We have a prior distribution for θ with density

$$f^{(0)}(\theta) = \begin{cases} k_0\{\theta^2(1-\theta) + \theta(1-\theta)^2\} & (0 < \theta < 1) \\ 0 & (\text{otherwise}) \end{cases} .$$

(a) i. Find the value of k_0.
ii. Find the prior mean of θ.
iii. Find the prior standard deviation of θ.

(b) We observe x, an observation from the binomial(n, θ) distribution.
i. Find the likelihood.
ii. Find the posterior density of θ.
iii. Find the posterior mean of θ.

3. We are interested in the parameter θ, of a binomial(n, θ) distribution. We have a prior distribution for θ with density

$$f^{(0)}(\theta) = \begin{cases} k_0\theta^2(1-\theta)^3 & (0 < \theta < 1) \\ 0 & (\text{otherwise}) \end{cases} .$$

(a) i. Find the value of k_0.
ii. Find the prior mean of θ.
iii. Find the prior standard deviation of θ.

(b) We observe x, an observation from the binomial(n, θ) distribution.
i. Find the likelihood.
ii. Find the posterior density of θ.
iii. Find the posterior mean of θ.

4. In a manufacturing process packages are made to a nominal weight of 1kg. All underweight packages are rejected but the remaining packages may be slightly overweight. It is believed that the excess weight X, in g, has a continuous uniform distribution on $(0, \theta)$ but the value of θ is unknown. Our prior density for θ is

$$f^{(0)}(\theta) = \begin{cases} 0 & (\theta < 0) \\ k_0/100 & (0 \le \theta < 10) \\ k_0\theta^{-2} & (10 \le \theta < \infty) \end{cases} .$$

(a) i. Find the value of k_0.
ii. Find the prior median of θ.

(b) We observe 10 packages and their excess weights, in g, are as follows.

3.8 2.1 4.9 1.8 1.7 2.1 1.4 3.6 4.1 0.8

Assume that these are independent observations, given θ.
i. Find the likelihood.
ii. Find a function $h(\theta)$ such that the posterior density of θ is $f^{(1)}(\theta) = k_1 h(\theta)$, where k_1 is a constant.
iii. Evaluate the constant k_1. *(Note that it is a very large number but you should be able to do the evaluation using a calculator).*

5. Repeat the analysis of the Chester Road example in section 6.3, using the same likelihood but with the following prior density.

$$f^{(0)}(\lambda) = \begin{cases} k_0[1+(8\lambda)^2]^{-1} & (0 < \lambda < \infty) \\ 0 & (\text{otherwise}) \end{cases}$$

   (a) Find the value of k_0.

   (b) Use numerical methods in R to do the following.

      i. Find the posterior density and plot a graph showing both the prior and posterior densities.

      ii. Find the posterior mean and standard deviation.

   Note: For the numerical calculations and the plot in part (b) I suggest that you use a range $0.0 \leq \lambda \leq 0.2$. When plotting the graph, it is easiest to plot the posterior first as this will determine the length of the vertical axis. The value of k_0 can be found analytically. If you do use numerical integration to find it, you will need a much wider range of values of λ.

6. We are interested in the parameter λ of a Poisson(λ) distribution. We have a prior distribution for λ with density

$$f^{(0)}(\lambda) = \begin{cases} 0 & (\lambda < 0) \\ k_0\lambda^3 e^{-\lambda} & (\lambda \geq 0) \end{cases} .$$

   (a) i. Find the value of k_0.

      ii. Find the prior mean of λ.

      iii. Find the prior standard deviation of λ.

   (b) We observe $x_1, \ldots, x_n$ which are independent observations from the Poisson(λ) distribution.

      i. Find the likelihood function.

      ii. Find the posterior density of λ.

      iii. Find the posterior mean of λ.

7. In a fruit packaging factory apples are examined to see whether they are blemished. A sample of n apples is examined and, given the value of a parameter θ, representing the proportion of aples which are blemished, we regard x, the number of blemished apples in the sample, as an obervation from the binomial($n,\ \theta$) distribution. The value of θ is unknown.

   Our prior density for θ is

$$f^{(0)}(\theta) = \begin{cases} k_0(20\theta(1-\theta)^3 + 1) & (0 \leq \theta \leq 1) \\ 0 & (\text{otherwise}) \end{cases} .$$

   (a) i. Show that, for $0 \leq \theta \leq 1$, the prior density can be written as

$$f^{(0)}(\theta) = \frac{1}{2}\left\{ \frac{\Gamma(6)}{\Gamma(2)\Gamma(4)}\theta^{2-1}(1-\theta)^{4-1} + \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}\theta^{1-1}(1-\theta)^{1-1} \right\}.$$

      ii. Find the prior mean of θ.

      iii. Find the prior standard deviation of θ.

   (b) We observe $n = 10$ apples and $x = 4$.

      i. Find the likelihood function.

      ii. Find the posterior density of θ.

      iii. Find the posterior mean of θ.

      iv. Use R to plot a graph showing both the prior and posterior densities of θ. (*Hint: It is easier to get the vertical axis right if you plot the posterior density and then superimpose the prior density, rather than the other way round.*)

6.7 Homework 2

Solutions to Questions **6, 7** *of Problems 2 are to be submitted in the Homework Letterbox no later than 4.00pm on Monday February 23rd.*