1 Elementary Bayes

Once we accept the notion of subjective probability, in which probability is interpreted as a "measure of belief", the Bayesian method follows naturally from a mechanistic application of the laws of probability.

If x and y are random quantities, then

- $p(x, y)$ – is the joint density function of x and y, that describes the probability that combinations of x and y will occur together,
- $p(x \mid y)$ – is the conditional density of x given y, and describes the distribution of x for a given value y, and
- $p(x)$ – is the marginal density of x, and describes the distribution of x irrespective of y.

Given the joint density $p(x, y)$, the marginal density $p(x)$ is determined b averaging over y

$$p(x) = \int p(x, y)dy.$$

The conditional density is defined as

$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

which implies that $p(x, y) = p(x \mid y)p(y)$. More generally

$$p(x, y, z, \ldots, a, b, c, \ldots) = p(x, y, z, \ldots \mid a, b, c, \ldots)p(a, b, c, \ldots).$$

Bayes rule follows from this decomposition of the joint density, as

$$p(x \mid y)p(y) = p(x, y) = p(y \mid x)p(x)$$

so

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} = \frac{p(y \mid x)p(x)}{\int p(x, y)dx}.$$

2 Binomial Responses

In this section we introduce the Binomial distribution and consider a range of common models for Binomial responses.

2.1 Binomial Distribution

A sequence of Bernoulli trials is a sequence of random experiments in which

- the outcome in each experiment is binary – either an event A occurs or fails to occur,

- the trials are independent – the outcome of any trial is not affected by the outcome of any other trial, and
- the probability of observing an A in a single trial is the same for all trials.

Commonly the occurrence of A is described as a success and the absence of A as a failure.

The archetypal example is repeatedly tossing a coin. Each toss is a trial, the trials are independent as the outcome of one toss in no way influences any other, and the probability of "heads" in any single trial remains constant.

The Binomial distribution describes the number of "successes" Y from n Bernoulli trials. We write

$$Y \sim \text{Bin}(n, \pi)$$

where Y is the number of successes, n the number of trials, and π the probability of a success in any single trial. The Binomial distribution has probability mass function

$$P(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

Here the term π^y is the probability of the occurrence of y successes, the term $(1 - \pi)^{n-y}$ is the probability of the occurrence of $n - y$ failures, while the term $\binom{n}{y}$ is the number of sequences successes and failures that would lead to a total of y successes. For example, if we toss a coin twice, then only the sequence HH can yield two heads, but a single head can be obtained as either HT or TH.

Suppose a non-contagious disease occurs in the general population with an incidence of 0.03. If Y denotes the number of cases we observe in a random sample of 100 people, then we may reasonably assume that

$$Y \sim \text{Bin}(100, 0.03)$$

and the probability of observing 5 cases in the random sample of 100 people is

$$\begin{aligned} P(Y = 5) &= \binom{100}{5} 0.03^5 (1 - 0.03)^{100-5} \\ &= 0.1013 \end{aligned}$$

2.2 Estimating a Binomial Probability

Once a probability model is assumed we can determine the probability of observing particular data given known values of model parameters. Bayes' rule allows us to invert this relation, to derive estimates of unknown model parameters given an observed data set.

Given the likelihood $p(y|\theta)$ – the probability of occurrence of the observed data y for given values of the parameter θ, Bayes' rule relates the posterior distribution $p(\theta|y)$ – the distribution of the unknown parameters given the observed data, to the prior distribution $p(\theta)$ – the distribution assumed for the

parameters in the absence of data,

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}.$$

Suppose we observe y successes from n Bernoulli trials, and we wish to estimate π, the probability of a success in a single trial. Suppose we adopt a Beta prior for π

$$\pi \sim \text{Beta}(a, b)$$

for some known a and b, so that

$$p(\pi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1}.$$

The response is Binomial,

$$y \mid \pi \sim \text{Bin}(n, \pi)$$

and so the likelihood is

$$p(y \mid \pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}.$$

The posterior density can then be computed mechanically by Bayes' rule

$$\begin{aligned}
p(\pi|y) &= \frac{p(y|\pi)p(\pi)}{\int_0^1 p(y|\pi)p(\pi)d\pi} \\
&= \frac{\binom{n}{y}\Gamma(a+b)\Gamma(a)^{-1}\Gamma(b)^{-1}\pi^{y+a-1}(1-\pi)^{n-y+b-1}}{\int_0^1 \binom{n}{y}\Gamma(a+b)\Gamma(a)^{-1}\Gamma(b)^{-1}\pi^{y+a-1}(1-\pi)^{n-y+b-1}d\pi} \\
&= \frac{\pi^{y+a-1}(1-\pi)^{n-y+b-1}}{\int_0^1 \pi^{y+a-1}(1-\pi)^{n-y+b-1}d\pi} \\
&= \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}\pi^{y+a-1}(1-\pi)^{n-y+b-1}
\end{aligned}$$

We can recognize this as the density function of a Beta distribution, and deduce that

$$\pi \mid y \sim \text{Beta}(y+a, n-y+b).$$

Deriving the posterior through the straightforward applications of Bayes' rule approach requires the evaluation of an unpleasant integral. A much simpler alternative is to recognize that the integral in the denominator of Bayes' rule is a normalizing constant, and

$$\begin{aligned}
p(\pi \mid y) &\propto p(y \mid \pi)p(\pi) \\
&\propto \left[\binom{n}{y}\pi^y(1-\pi)^{n-y}\right] \times \left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1}\right] \\
&\propto \pi^{y+a-1}(1-\pi)^{n-y+b-1}.
\end{aligned}$$

The only distribution with this functional form is the Beta, so again we deduce

$$\pi \mid y \sim \text{Beta}(y + a, n - y + b).$$

Continuing the example from the previous section, suppose the incidence of the disease is unknown, but in a random sample of 100 people we detect 5 cases. If we take as prior

$$\pi \sim \text{Beta}(1, 1),$$

the posterior is

$$\pi \sim \text{Beta}(6, 96).$$

and is shown in Figure 1. The dotted lines delimit a 95% credible interval for π – that is, 2.5% of the probability mass lies below the line on the left, and 2.5% of the probability mass lies above the line on the right. This is why it is essential to know the full posterior, not just its kernel – the distribution must be normalized to allow us to make probabilistic statements about π.

2.2.1 WinBUGS

We can simulate from the posterior with WinBUGS, and hence determine estimates of the posterior mean and credible intervals.

In WinBUGS the model is represented as

```
model{
  ## Prior
  pi ~ dbeta(1,1)
  ## Likelihood
  y ~ dbin(pi,n)
}

## Data
list(n=100,y=5)
```

This program effectively mirrors the probability statements

$$\pi \sim \text{Beta}(a, b)$$
$$y \sim \text{Bin}(n, \pi)$$

although annoyingly, WinBUGS reverses the conventional order of parameters to the Binomial distribution.

2.2.2 Multiple Observations

Suppose instead we had collected a data from a random sample of 100 people in each of 3 different cities, and y_1 cases were observed in the first, y_2 in the second and y_3 in the third, and let us assume the incidence of the disease is the same in each city. Assuming the incidence is the same in each city

$$y_i \sim \text{Bin}(n_i, \pi)$$

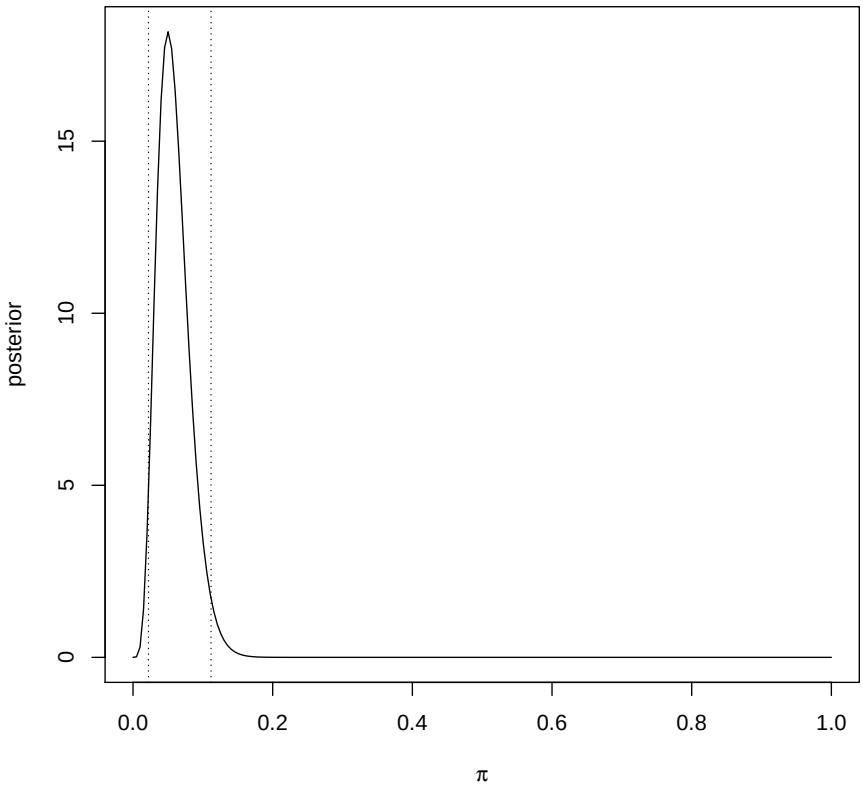

Figure 1: Posterior distribution for π when 5 cases are observed in a random sample of 100 people.

Because each sample is independent, the likelihood is simply the product of the likelihoods for each sample

$$\begin{aligned} p(y_1, y_2, y_3|\pi) &= p(y_1|\pi)p(y_1|\pi)p(y_1|\pi) \\ &= \prod_{i=1}^{3} \binom{n_i}{y_i} \pi^{y_i}(1-\pi)^{n_i - y_i} \\ &\propto \pi^{\sum_{i=1}^{3} y_i}(1-\pi)^{\sum_{i=1}^{3}(n_i - y_i)} \end{aligned}$$

If we again adopt a Beta prior,

$$\pi \sim \text{Beta}(1, 1),$$

we deduce that

$$\pi|y_1, y_2, y_3 \sim \text{Beta}(a + \sum_{i=1}^{3} y_i, b + \sum_{i=1}^{3}(n_i - y_i)).$$

Note that the posterior does not depend directly on the individual y_i and n_i, only their sums. The sums S_n and S_y are said to be *sufficient statistics* for π, in that they contain all the relevant information for estimating π.

Figure 2 shows the posterior for $n_1 = n_2 = n_3 = 100$, and $y_1 = 5$, $y_2 = 8$ and $y_3 = 1$ and $a = b = 1$

In this case, the WinBUGS program takes the form

```
model {
  ## Prior
  pi ~ dbeta(1,1)
  ## Likelihood
  for(i in 1:3) {
    y[i] ~ dbin(pi,n[i])
  }
}
##Data
list(y=c(5,8,1),n=c(100,100,100))
```

2.3 Binomial Regression

Consider the case where we record not only a Binomial distributed response y, but also explanatory variables $x_1, x_2, \ldots, x_m$, and we wish to model the role the explanatory variables play in determining the probability of a success π in each trial. For example, we may be monitoring incursions of an invasive species at sites around the state, and our explanatory variables may be habitat descriptors at the sites.

The Binomial generalized linear model assumes that the response y is Binomially distributed

$$y_i \sim \text{Bin}(n_i, \pi_i)$$

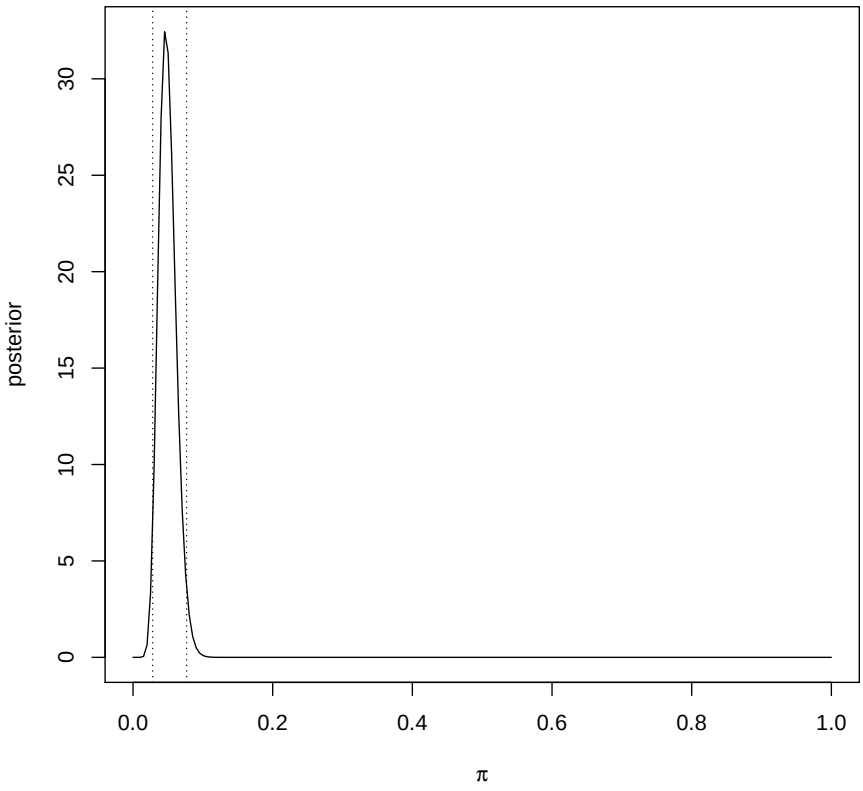

Figure 2: Posterior distribution for π when 5, 8 and 1 cases are observed in three random samples of 100 people.

with a probability π of success in an an individual trial is related to a linear combination of predictors $x_1, x_2, \ldots, x_m$ through a monotonic link function l

$$l(\pi_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi}.$$

The three common choices of link function are:

- the logit link $f(\pi) = \log \frac{\pi}{1-\pi}$,
- the probit link $f(\pi) = \phi^{-1}(\pi)$ where ϕ is the standard Normal distribution function, and
- the complementary log-log link $f(\pi) = \log(-\log(1-\pi))$.

Although we can readily determine an expression for the likelihood for the model, in practice it is not possible to evaluate the multidimensional integral

$$\int p(y|\beta)p(\beta)d\beta$$

that occurs in the denominator of Bayes' rule, and we must resort to simulation techniques to fit this model.

For the same reason, there is no closed form conjugate prior for this problem, and so common choices for the prior are to adopt either an improper uniform prior

$$\beta_i \sim \mathrm{U}(-\infty, \infty)$$

or a diffuse Normal prior

$$\beta_i \sim \mathrm{N}(\mu, \sigma^2)$$

where σ^2 is large.

In this case, Monte Carlo Markov Chain (MCMC) methods are essential. A WinBUGS program for this problem that assumes diffuse Normal priors for the β_i, a logit link function and two covariates takes the general form

```
model {
  ## Prior
  for(j in 1:3) {
    beta[j] ~ dnorm(0,0.001)
  }
  ## Likelihood
  for(i in 1:n) {
    logit(pi[i]) <- beta[1] + beta[2]*x1[i] + beta[3]*x2[i]
    y[i] ~ dbin(pi[i],n[i])
  }
}
##Data
list(y=c(...),
     n=c(...),
     x1=c(...),
     x2=c(...))
```

Here there are few subtle differences from our mathematical notation. Vectors in WinBUGS are indexed from 1 so `beta[1]` must play the role of β_0. Also, WinBUGS parametrizes the Normal distribution in terms of its mean μ and precision $\tau = \sigma^{-2}$, so a diffuse prior is one with τ small.

We cannot adopt the improper uniform prior here – WinBUGS syntax simply does not allow it.

2.4 Random Effects

One strength of the MCMC approach is that it allows us to fit models that are extremely difficult to fit with classical methods. A common example is models containing random effects.

The standard Binomial regression model

$$\begin{aligned} y_i &\sim \text{Bin}(n_i, \pi_i) \\ l(\pi_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} \end{aligned}$$

assumes that the only source of random variability is associated with the Binomial nature of the responses. Under this model two responses with precisely the same values for the explanatory covariates should have identical probability of success π, and all of the variability in the response arises from the Binomial distribution. Often this is too naive; there may be sources of variability in π that cannot be fully explained by the explanatory covariates. These unidentified sources of additional variability can be modelled by the inclusion of additional random terms or random effects in the linear predictor.

For example, consider the following experiment. An experiment was conducted to investigate the carcinogenic effects of a toxic by-product of a mold that infects cottonseed and grains. Twenty tanks were of rainbow trout embryos were exposed to one of five doses of aflatoxicol for one hour. After one year, the fish were dissected and the number of fish in each tank with liver tumours recorded.

In this case the response is the number of fish with liver tumours recorded in each tank, and the explanatory variable is the level of exposure to aflatoxicol. Yet it is somewhat naive to assume that two tanks that receive the same exposure are actually identical replicates – there may be subtle tank to tank differences that are not reflected simply by the aflatoxicol dose.

We could model this by incorporating an additional random component into the linear predictor,

$$\begin{aligned} \text{Tumour}_{ij} &\sim \text{Bin}(n_{ij}, \pi_{ij}) \\ l(\pi_{ij}) &= \text{Dose}_i + \text{Tank}_{ij} \\ \text{Tank}_{ij} &\sim \text{N}(0, \tau_T) \end{aligned}$$

where Tumour_{ij} denotes the number of tumourous fish in the j-th tank receiving the dose i, Dose_i denotes the effect of dose level i and Tank_{ij} denotes the random

tank effect. Here we have assumed that the tank effects are Normally distributed with precision τ_T^2.

In WinBUGS, our model would take the form

```
model {
  ## loop over doses
  for(i in 1:D) {
    ## loop over tanks
    for(j in 1:T) {
      ## logit model for tumour incidence
      y[i,j] ~ dbin(p[i,j],n[i,j])
      logit(p[i,j]) <- dose[i]+tank[i,j]
      ## distribution of random tank effects
      tank[i,j] ~ dnorm(0,tau.tank)
    }
  }

  ## prior for treatment means
  for(i in 1:D) {
    dose[i] ~ dnorm(0,0.0001)
  }
  ## prior for precision of tank effects
  tau.tank ~ dgamma(0.001,0.001)
}
```

3 Poisson Responses

In this section we introduce the Poisson process and consider a range of common models for Poisson responses.

3.1 Poisson Distribution

A Poisson point process is a stream of random events in which

- the rate at which events occur is constant, and
- the occurrence of any two events is independent.

The Poisson distribution describes the number of events Y of a Poisson process occurring in a unit interval. We write

$$Y \sim \text{Poisson}(\lambda)$$

where Y is the number of events and λ the rate at which events occur. The Poisson distribution has probability mass function

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

One approach to deriving the Poisson distribution is to view it as a limiting form of the Binomial. If $\pi \to 0$ and $n \to \infty$ in such a way that $n\pi = \lambda$, then

$$\binom{n}{y}\pi^y(1-\pi)^{n-y} \to \frac{e^{-\lambda}\lambda^y}{y!}.$$

Suppose the number of major breakdowns on a network occur according to a Poisson process with a rate of 2.5 breakdowns per week. Then the probability of observing 5 major breakdowns per week is

$$\begin{aligned} P(Y=5) &= \frac{e^{-2.5}2.5^5}{5!} \\ &= 0.0668 \end{aligned}$$

3.2 Estimating a Poisson Rate

Suppose that in a unit interval, we observe y events from a Poisson process, and we wish to estimate the Poisson rate λ. Suppose we adopt a Gamma prior for λ

$$\lambda \sim \text{Gamma}(a, b)$$

for some a and b, so that

$$p(\lambda) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}.$$

The response is Poisson

$$y \mid \lambda \sim \text{Poisson}(\lambda)$$

and so the likelihood is

$$p(y \mid \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

By Bayes' rule

$$\begin{aligned} p(\lambda \mid y) &\propto p(y \mid \lambda)p(\lambda) \\ &\propto \left[\frac{e^{-\lambda}\lambda^y}{y!}\right] \times \left[\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}\right] \\ &\propto e^{-(b+1)\lambda}\lambda^{y+a-1}. \end{aligned}$$

We recognize as the functional form of the Gamma distribution, and so we deduce the posterior

$$\lambda \mid y \sim \text{Gamma}(y+a, b+1).$$

This shows that the Gamma distribution is a conjugate prior for the Poisson distribution.

Extending the example from the previous section, suppose instead the rate of network breakdowns is unknown but in a single week we observe $y = 5$ breakdowns. Suppose we take as prior the Gamma distribution

$$\lambda \sim \text{Gamma}(a, b),$$

which gives prior mean $\text{E}(\lambda) = 1$ and prior variance $\text{Var}(\lambda) = 100$. The posterior is

$$\lambda \mid y \sim \text{Gamma}(5.01, 1.01),$$

giving a posterior mean $\text{E}(\lambda \mid y) \approx 4.96$ and posterior variance $\text{Var}(\lambda \mid y) \approx 4.91$. The posterior is shown in Figure 3. The dotted lines delimit a 95% credible interval for λ.

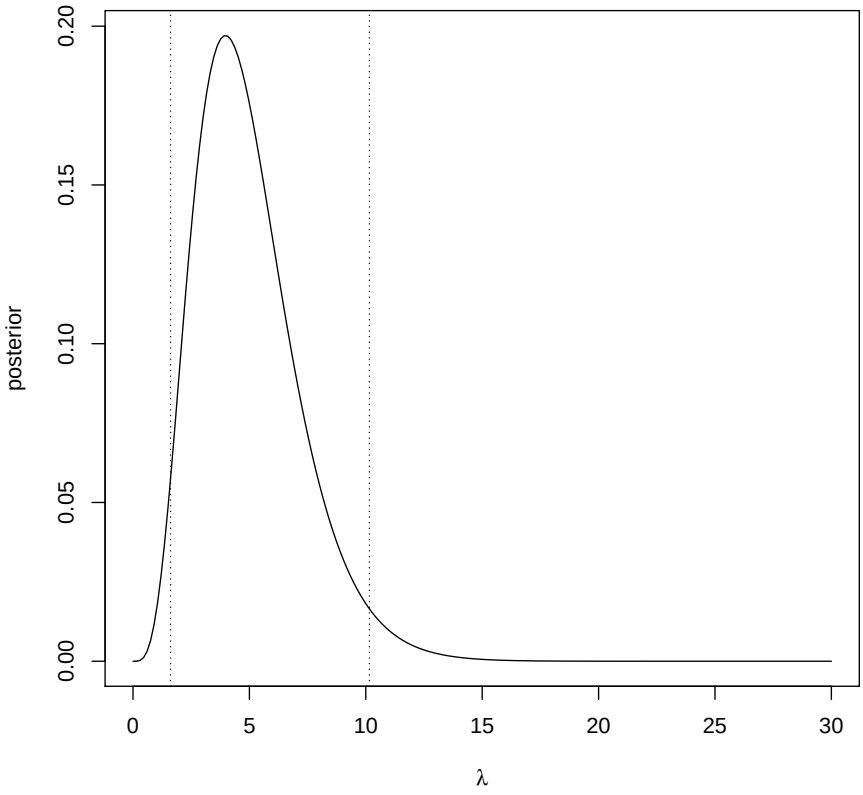

Figure 3: Posterior distribution for λ when 5 breakdowns are observed in a week.

3.2.1 WinBUGS

To represent this model in WinBUGS

```
model{
  ## Prior
  lambda ~ dgamma(0.01,0.01)
  ## Likelihood
  y ~ dpois(lambda)
}

## Data
list(y=5)
```

This program effectively mirrors the probability statements

$$\begin{aligned}\lambda &\sim \text{Gamma}(0.01, 0.01)\\ y &\sim \text{Poisson}(\lambda).\end{aligned}$$

3.2.2 Multiple Observations

Suppose instead we had collected a data over 3 weeks, and observed y_1 breakdowns in the first week, y_2 in the second and y_3 in the third, so that

$$y_i \sim \text{Poisson}(\lambda)$$

Because each sample is independent, the likelihood is simply the product of the likelihoods for each sample

$$\begin{aligned}p(y_1, y_2, y_3|\lambda) &= p(y_1|\lambda)p(y_1|\lambda)p(y_1|\lambda)\\ &= \prod_{i=1}^{3} e^{-\lambda}\lambda^{y_i}/y_i!\\ &= \frac{e^{-3\lambda}\lambda^{\sum_{i=1}^{3} y_i}}{y_1!y_2!y_3!}.\end{aligned}$$

By Bayes' rule

$$\begin{aligned}p(\lambda \mid y_1, y_2, y_3) &\propto p(y_1, y_2, y_3 \mid \lambda)p(\lambda)\\ &\propto e^{-(b+3)\lambda}\lambda^{a-1+\sum_{i=1}^{3} y_i}\end{aligned}$$

which we again recognize as the kernel of a Gamma distribution

$$\lambda \mid y_1, y_2, y_3 \sim \text{Gamma}(a + \sum_{i=1}^{3} y_i, b + 3).$$

Figure 4 shows the posterior for $y_1 = 5$, $y_2 = 8$ and $y_3 = 1$.
In this case, the WinBUGS program takes the form

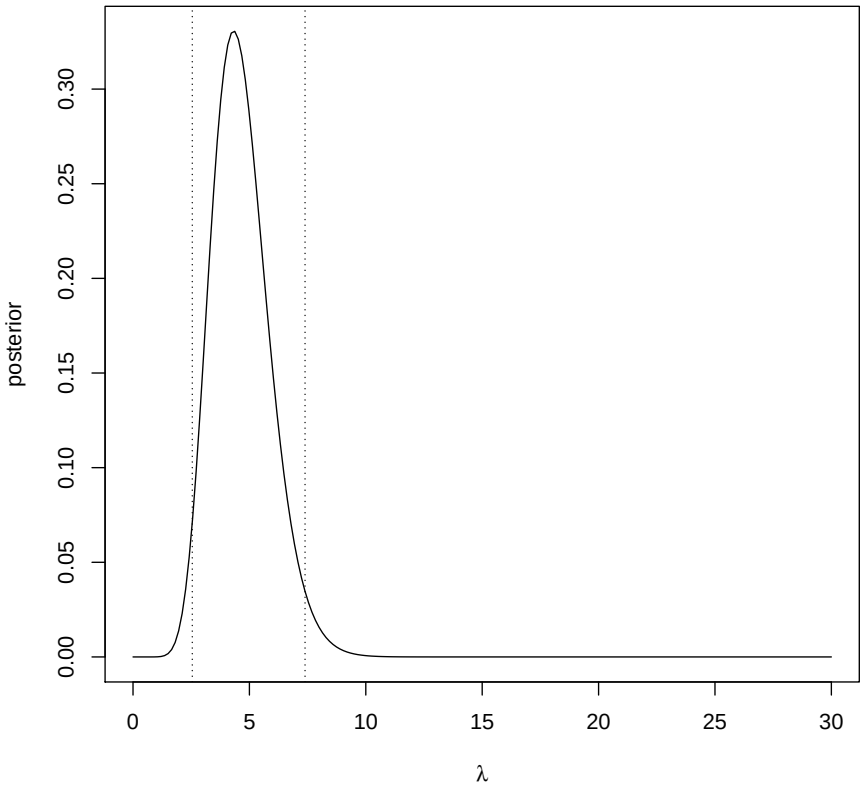

Figure 4: Posterior distribution for λ when 5, 8 and 1 breakdowns are observed over three weeks.

```
model {
  ## Prior
  lambda ~ dgamma(0.001,0.001)
  ## Likelihood
  for(i in 1:3) {
    y[i] ~ dpois(lambda)
  }
}
##Data
list(y=c(5,8,1))
```

3.3 Poisson Regression

The standard Poisson regression model takes the from

$$\begin{aligned} y_i &\sim \text{Poisson}(\lambda_i) \\ l(\lambda_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} \end{aligned}$$

where y_i denotes the i-th observation of the response, and x_{ji} denote the i-th observation of the j-th explanatory variable. For Poisson regression it is typically to use a log link function, $l(\lambda_i) = \log(\lambda_i)$.

As for Binomial regression models, some form of Monte Carlo Markov Chain (MCMC) method is essential to a Bayesian treatment of the Poisson regression model.

If we assume diffuse Normal priors for the β_i, a log link function and if $m = 2$, a WinBUGS program for the Poisson regression problem takes the form

```
model {
  ## Prior
  for(j in 1:3) {
    beta[j] ~ dnorm(0,0.001)
  }
  ## Likelihood
  for(i in 1:n) {
    log(lambda[i]) <- beta[1] + beta[2]*x1[i] + beta[3]*x2[i]
    y[i] ~ dpois(lambda[i])
  }
}
##Data
list(y=c(...),
     n=c(...),
     x1=c(...),
     x2=c(...))
```

Random effects may be added to the model in much the same way as for a Binomial regression model (Section 2.4).

4 Normal Responses

In this section we consider simple one sample and regression problems for Normally distributed data.

We assume that the reader is familiar with the basic properties of the Normal distribution, but point out that in the Bayesian literature, it is common to parametrize the Normal distribution in terms of its mean μ and precision $\tau = \sigma^{-2}$, the inverse of the variance.

4.1 Estimating a Normal Mean

Suppose we observe n independent observations $y_1, y_2, \ldots, y_n$ from a Normal distribution with mean μ and precision τ

$$y_i \sim \mathrm{N}(\mu, \tau)$$

and we wish to estimate μ. This example differs from the others we seen in that both μ and τ are unknown. But our focus is μ, we have little interest in τ, we say it is a "nuisance" parameter.

We adopt a Normal prior for μ and an independent Gamma prior for τ

$$\begin{aligned} \mu &\sim \mathrm{N}(\mu_0, \tau_0) \\ \tau &\sim \mathrm{Gamma}(a, b) \end{aligned}$$

for some μ_0, τ_0, a and b, so that

$$\begin{aligned} p(\mu, \tau) &= p(\mu)p(\tau) \\ &= \left[\left(\frac{\tau_0}{2\pi} \right)^{\frac{1}{2}} e^{-\tau_0(\mu-\mu_0)^2/2} \right] \left[\frac{b^a}{\Gamma(a)} \tau^{a-1} e^{-b\tau} \right]. \end{aligned}$$

The y_i are Normally distributed, so the likelihood takes the form

$$\begin{aligned} p(y_1, \ldots, y_n \mid \mu, \tau) &= \prod_{i=1}^{n} \left(\frac{\tau}{2\pi} \right)^{\frac{1}{2}} e^{-\tau(y_i-\mu)^2/2} \\ &= \left(\frac{\tau}{2\pi} \right)^{\frac{n}{2}} e^{-\tau \sum_{i=1}^{n} (y_i-\mu)^2/2} \end{aligned}$$

and by Bayes rule

$$\begin{aligned} &p(\mu, \tau \mid y_1, \ldots, y_n) \\ &\quad \propto p(y_1, \ldots, y_n \mid \mu, \tau) p(\mu, \tau) \\ &\quad \propto \tau^{n/2+a-1} e^{-\left[\tau_0(\mu-\mu_0)^2/2 + \tau \sum_{i=1}^{n} (y_i-\mu)^2/2\right]} \end{aligned}$$

Ideally we would like to normalize this expression to obtain $p(\mu, \tau \mid y_1, \ldots, y_n)$, the joint posterior of μ and τ, then average over τ to obtain $p(\mu \mid y_1, \ldots, y_n)$ the posterior marginal distribution of μ. But even for this simple example the mathematics is formidable, and not particularly enlightening – as one might expect, after integrating out τ the posterior marginal distribution of μ is recognized as a t-distribution.

4.2 Gibbs Sampling

To simulate from the posterior by Gibbs sampling, we require the conditional distributions of μ and τ.

By definition

$$p(\mu, \tau \mid y_1, \ldots, y_n) = p(\mu \mid \tau, y_1, \ldots, y_n) p(\tau \mid y_1, \ldots, y_n).$$

But on the right hand side, only the conditional distribution $p(\mu \mid \tau, y_1, \ldots, y_n)$ is a function of μ. So to derive $p(\mu \mid \tau, y_1, \ldots, y_n)$, we focus on $p(\mu, \tau \mid y_1, \ldots, y_n)$ as a function of μ alone, ignoring any dependence on τ. If we expand and only retain terms that depend on μ, we find

$$p(\mu, \tau \mid y_1, \ldots, y_n) \propto e^{-\left[\tau_0 (\mu - \mu_0)^2 / 2 + \tau \sum_{i=1}^n (y_i - \mu)^2 / 2\right]},$$

which can be further simplified to show that

$$p(\mu \mid \tau, y_1, \ldots, y_n) \propto p(\mu, \tau \mid y_1, \ldots, y_n) \propto e^{\tau_1 (\mu - \mu_1)^2 / 2}$$

where

$$\begin{aligned} \mu_1 &= \frac{\tau_0 \mu_0 + n\tau \bar{y}}{\tau_0 + n\tau} \\ \tau_1 &= \tau_0 + n\tau \\ \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i. \end{aligned}$$

We recognise this as the functional form of a Normal density, and so deduce that

$$\mu \mid \tau, y_1, \ldots, y_n \sim \mathrm{N}(\mu_1, \tau_1).$$

This implies that if τ is known, μ is a precision weighted average of the the prior mean μ_0 and the sample mean $\bar{y}$.

The conditional for τ is derived in a similar fashion. If we expand and only retain terms that depend on τ, we find

$$p(\mu, \tau \mid y_1, \ldots, y_n) \propto \tau^{n/2 + a - 1} e^{-\frac{1}{2} \tau \sum_{i=1}^n (y_i - \mu)^2}.$$

This expression can be recognized as the density of a Gamma distribution

$$\tau \mid \mu, y_1, \ldots, y_n \sim \mathrm{Gamma}\left(a + n/2, b + S/2\right).$$

where

$$S = \sum_{i=1}^n (y_i - \mu)^2.$$

This conditional distribution has mean

$$\mathrm{E}[\tau \mid \mu, y_1, \ldots, y_n] = \frac{n + 2a}{S + 2b}$$

which can be seen to be a compromise between the inverse of the usual "sum of squares" estimator of variance and the prior.

To Gibbs sample from the posterior $p(\mu, \tau \mid y_1, \ldots, y_n)$ we repeatedly sample from these two conditionals in turn. That is, given some initial point $(\mu^{(0)}, \tau^{(0)})$, we repeatedly sample first from

$$\mu^{(k+1)} \mid \tau^{(k)}, y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\tau_0\mu_0 + n\tau^{(k)}\bar{y}}{\tau_0 + n\tau^{(k)}}, \tau_0 + n\tau^{(k)}\right)$$

then from

$$\tau^{(k+1)} \mid \mu^{(k+1)}, y_1, \ldots, y_n \sim \mathrm{Gamma}\left(a + n/2, b + \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \mu^{(k+1)}\right)^2\right)$$

to generate a sequence of draws $(\mu^{(k)}, \tau^{(k)})$ from the posterior.