Chapter 9

Conjugate priors

9.1 The epidemiology of Lyme disease

Lyme disease is a tick-borne infectious disease spread by bacteria of species *Borrelia*, which are transmitted to ticks when they feed on animal hosts. Whilst fairly common in the US, this disease has recently begun to spread throughout Europe.

Imagine you are researching the occurrence of Lyme disease in the UK. As such, you begin by collecting samples of 10 ticks from fields and grasslands around Oxford, and counting the occurrence of the *Borrelia* bacteria.

Problem 9.1.1. You start by assuming that the occurrence of *Borrelia* bacteria in one tick is independent of that in other ticks. In this case, why is it reasonable to assume a binomial likelihood?

If we assume independence in disease between ticks (as well as assuming the underlying prevalence is the same across all surveyed terrains; i.e. identically-distributed), then because the data is discrete, and the sample size fixed $\implies$ **binomial** likelihood.

Problem 9.1.2. Suppose the number of *Borrelia*-positive ticks within each sample i is given by the random variable X_i, and that the underlying prevalence (amongst ticks) of this disease is θ. Write down the likelihood for sample i.

The likelihood is given by the binomial probability (through the equivalence principle):

$$\begin{aligned} L(\theta|X_i) &= Pr(X_i|\theta) \\ &= \binom{10}{X_i}\theta^{X_i}(1-\theta)^{10-X_i} \end{aligned}$$

Problem 9.1.3. Suppose that in your first sample of size 10 you find $X_1 = 1$ case of Borrelia. Graph the likelihood here and hence (by eye) determine the maximum likelihood estimate of θ.

In R this likelihood can be graphed using the following,

```
curve(dbinom(1, 10, x), 0, 1)
```

The likelihood is shown in Figure 9.1. The maximum likelihood estimate is at $\theta = 0.1$.

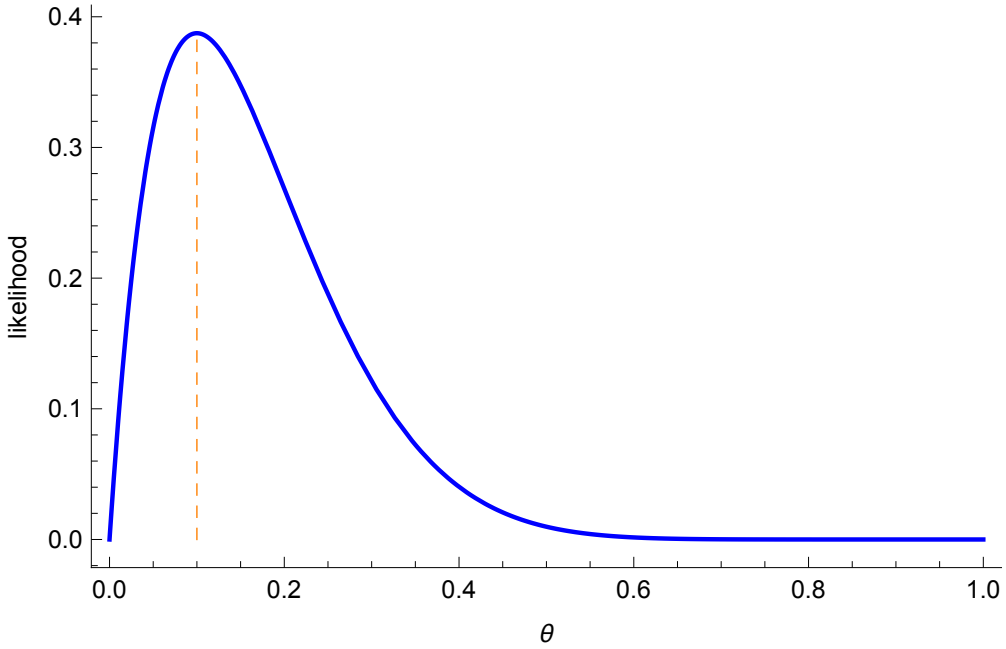

Figure 9.1: The likelihood for $X_1 = 1$ in the ticks example.

Problem 9.1.4. By numerical integration show that the area under the likelihood curve is about 0.09. Comment on this result.

In R this numerical integration can be carried out by the following,

```
integrate(function(x) dbinom(1, 10, x), 0, 1)
```

This is approximately $\frac{1}{11} \approx 0.09$. Therefore not a valid probability distribution!

Problem 9.1.5. Assuming that $\theta = 10\%$, graph the probability distribution (also known as the sampling distribution). Show that, in contrast to the likelihood, this distribution is a valid probability distribution.

This distribution can be graphed in R using,

```
lX <- seq(0, 10, 1)
plot(lX,sapply(lX, function(x) dbinom(x, 10, 0.1)),
     xlab="number of cases of bacteria out of 10", ylab="probability")
```

This is a discrete *probability* distribution shown in Figure 9.2. Since it is a discrete probability distribution we can check its validity by summing over all the probability masses,

```
lX <- seq(0, 10, 1)
sum(sapply(lX, function(x) dbinom(x, 10, 0.1))) == 1
```

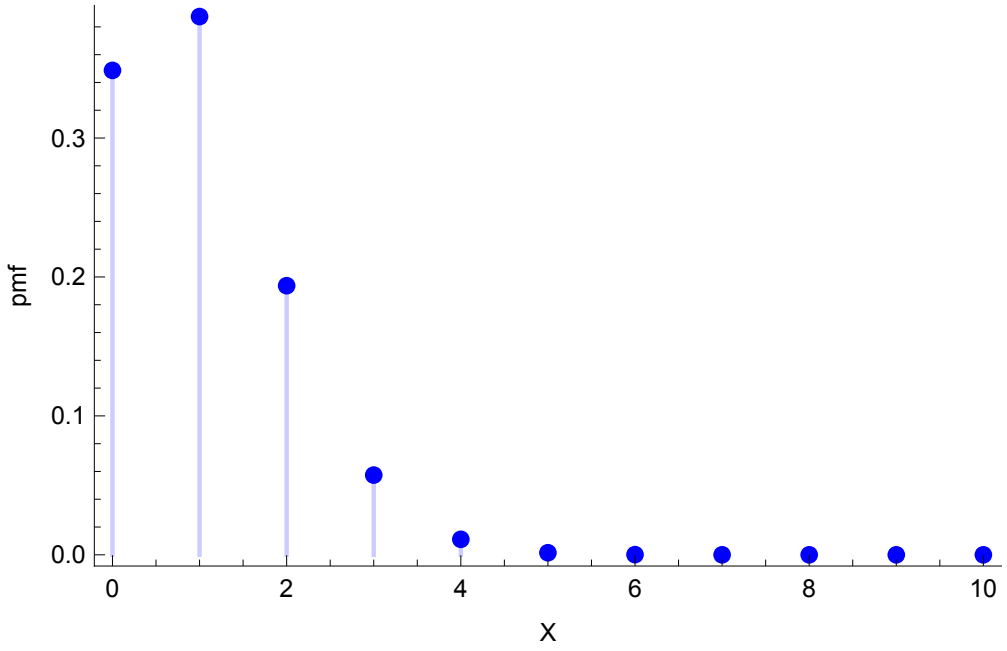

Figure 9.2: The sampling distribution for $\theta = 0.1$.

Problem 9.1.6. (Optional) Now assume that you do not know θ. Use calculus to show that the maximum likelihood estimator of the parameter, for a single sample of size 10 where we found X ticks with the disease is given by:

$$\hat{\theta} = \frac{X}{10} \tag{9.1}$$

(Hint: maximise the log-likelihood rather than the likelihood.)

We can write down the likelihood,

$$L(\theta|X) = \binom{10}{X}\theta^X(1-\theta)^{10-X}$$

Since the log is a monotonic transformation we can take the log of the likelihood, and maximise this instead. Taking the log we obtain,

$$l = log\ L(\theta|X) = \text{constants} + Xlog(\theta) + (10-X)log(1-\theta)$$

which we then differentiate to find the maximum,

$$\frac{\partial l}{\partial \theta} = \frac{X}{\hat{\theta}} - \frac{10 - X}{1 - \hat{\theta}} = 0 \tag{9.2}$$

which is obtained when $\hat{\theta} = \frac{X}{10}$.

Problem 9.1.7. A colleague mentions that a reasonable prior to use for θ is a $beta(a, b)$ distribution. Graph this for $a = 1$ and $b = 1$.

This is a continuous uniform distribution across (0,1), which can be obtained from the following R code,

```
curve(dbeta(x, 1, 1), 0, 1, xlab="theta", ylab="probability")
```

Problem 9.1.8. How does this distribution change as you vary a and b?

The mean is $\frac{a}{a+b}$. This can be obtained from R by doing,

```
?dbeta
```

and looking at the resultant help file. Therefore as $a \uparrow$ the mass of the distribution shifts to the right.

Problem 9.1.9. Prove that a $beta(a, b)$ prior is conjugate to the binomial likelihood, showing that the posterior distribution is given by a $beta(X + a, 10 - X + b)$ distribution.

- Likelihood:
$$X \sim \mathcal{B}(10, \theta) \implies p(X|\theta) \propto \theta^X (1 - \theta)^{10-X} \tag{9.3}$$
- For the prior assume a beta distribution (a reasonable choice if $\theta \in (0, 1)$):
$$\theta \sim beta(a, b) \implies p(\theta) \propto \theta^{a-1} (1 - \theta)^{b-1} \tag{9.4}$$
- Posterior:
$$\begin{aligned} p(\theta|X) &\propto p(X|\theta) \times p(\theta) \\ &\propto \theta^X (1 - \theta)^{10-X} \times \theta^{a-1} (1 - \theta)^{b-1} \\ &= \theta^{X+a-1} (1 - \theta)^{10-X+b-1} \end{aligned}$$

This has same θ-dependence as a $beta(X + a, 10 - X + b)$ density $\implies$ must be this distribution!

Problem 9.1.10. Graph the posterior for $a = 1$ and $b = 1$. How does the posterior distribution vary as you change the mean of the beta prior? (In both cases assume that $X = 1$.)

For $a = 1$ and $b = 1 \implies$ mean is $\frac{1+1}{10-1+1} = \frac{1}{5}$.

Problem 9.1.11. You now collect a larger dataset (encompassing the previous one) that has a sample size of 100 ticks in total; of which you find 7 carry *Borrelia*. Find and graph the new posterior using the conjugate prior rules for a $beta(1, 1)$ prior and binomial likelihood.

For $a = 1$ and $b = 1 \implies beta(1 + 7, 100 - 7 + 1)$ posterior, whose mean is $\frac{1+7}{100+2} = \frac{8}{102} \approx 0.078$. The posterior is shown in Figure 9.3, of which a similar curve can be obtained in R by doing the following,

```
curve(dbeta(x, 1 + 7, 100 - 7 + 1), 0, 1, xlab="theta", ylab="probability")
```

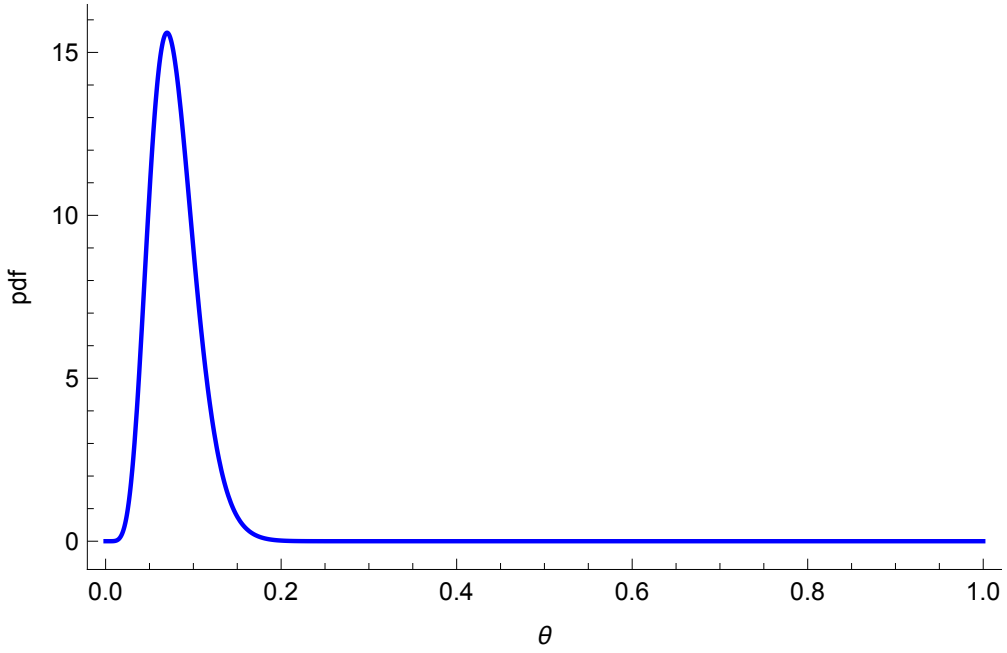

Figure 9.3: The posterior distribution for $X = 7$ out of a sample of 100 ticks.

Problem 9.1.12. You collect a second dataset of 100 ticks; this time finding that 4 carry the disease. Find and graph the new posterior (across both datasets) using the conjugate prior rules for a $beta(1, 1)$ prior and binomial likelihood. How does it compare to the previous one?

The new likelihood is the product of the two samples' likelihoods, and so we find a $beta(1 + 11, 200 - 11 + 1)$ posterior. This results in a narrower posterior (see Figure 9.4), which can similarly be produced in R using,

```
curve(dbeta(x, 1 + 11, 200 - 11 + 1), 0, 1, xlab="theta", ylab="probability")
```

Problem 9.1.13. Now we will use sampling to estimate the posterior predictive distribution for a sample size of 100, using the posterior distribution obtained from the entire sample of 200 ticks (11 of which were disease-positive). To do this we will first sample a random value of θ from the posterior: so $\theta_i \sim p(\theta|X)$. We then sample a random value of the data X by sampling from the

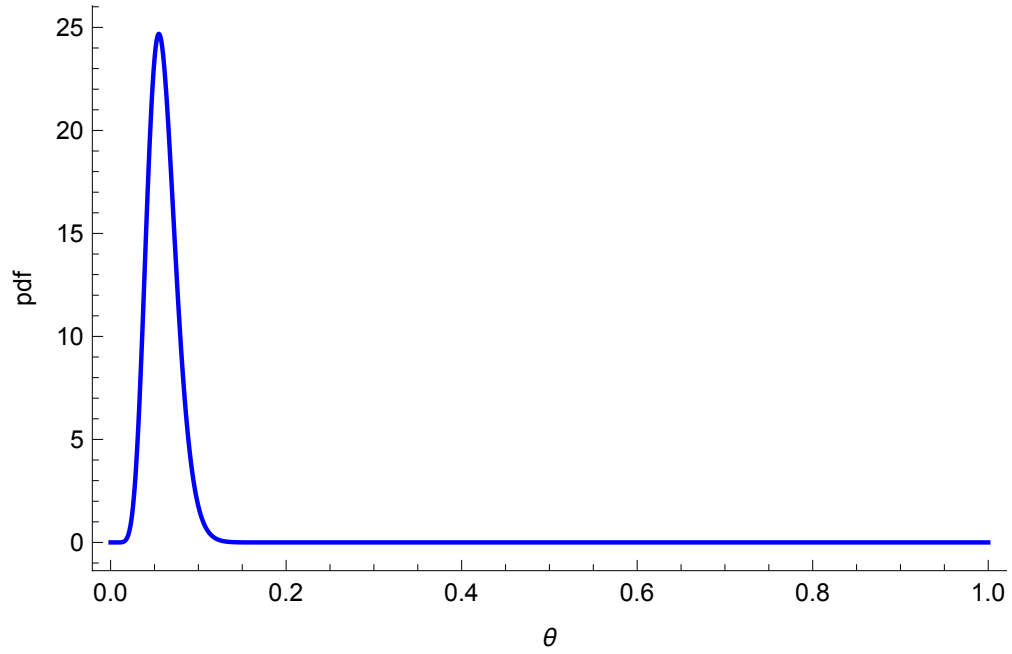

Figure 9.4: The posterior distribution for $X_1 = 7$ and $X_2 = 4$; each out of a sample of 100 ticks.

binomial sampling distribution $X_i \sim \mathcal{B}(100, \theta_i)$. We repeat this process a large number of times to obtain samples from this distribution. Follow the previous rules to produce 10,000 samples from the posterior predictive distribution, which we then graph using a histogram.

The posterior predictive distribution for a sample of 100 ticks is shown in Figure 9.5. I find the best way to do this is to create a function in R that does the above iteration,

```
fPosteriorPredictive <- function(aNumSamples){
  lX <- vector(length=aNumSamples)
  for(i in 1:aNumSamples){
    theta <- rbeta(1, 1 + 11, 200 - 11 + 1)
    X <- rbinom(1, 100, theta)
    lX[i] <- X
  }
  return(lX)
}
```

which we can then use to generate 10,000 posterior samples, then graph these using,

```
X <- fPosteriorPredictive(10000)
hist(X, breaks=seq(0, 100, 1), xlim = c(0, 20),
     xlab="number of disease-positive ticks")
```

Problem 9.1.14. Does our model fit the data?

Both the original data points are well contained within the posterior predictive distribution. Thus the model looks like a reasonable fit.

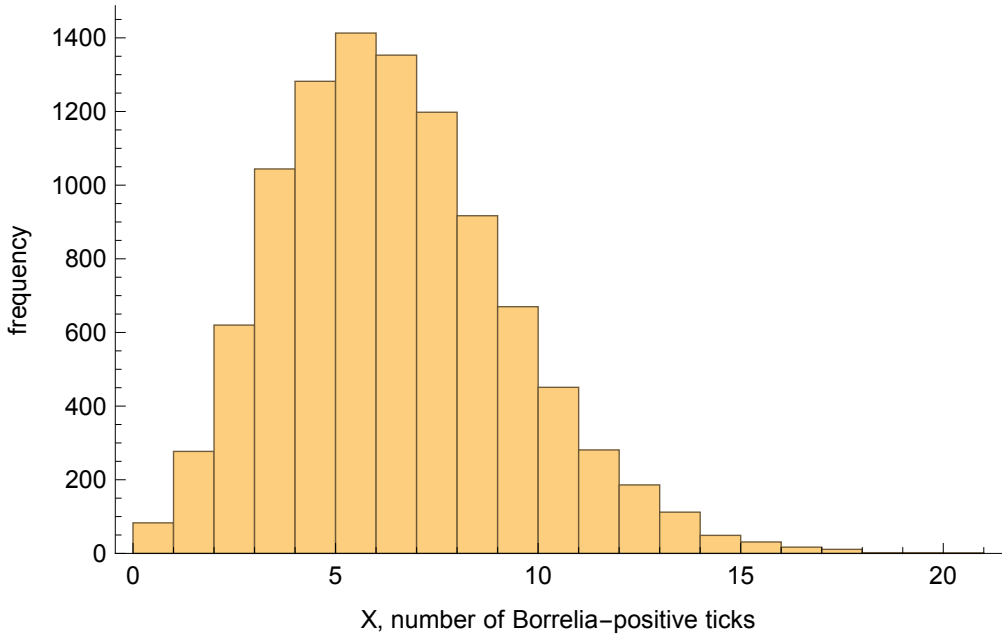

Figure 9.5: Samples from the posterior distribution predictive distribution for $X_1 = 7$ and $X_2 = 4$; for a sample size of 100 ticks.

Problem 9.1.15. Indicate whether you expect this model to hold across future sampling efforts.

Whilst it is a bit imprudent to comment on this, I would argue in this case that the assumption of **independence** of Borrelia amongst ticks is a bit suspect. In particular, the presence of one disease-positive tick makes it more likely that another - nearby - tick will catch the disease whilst blood-feeding. A more robust model might be preferable, for example the beta-binomial.

Problem 9.1.16. If we assume a uniform prior on θ, the probability that a randomly sampled tick carries Lyme disease, what is the shape of the prior for θ^2? (This is the probability that 2/2 ticks carry Lyme disease.)

Hint: do this either using Jacobians (hard-ish), or by sampling (easy-ish).

Assume a change of variables $y = g(x)$, how does the density change? We need the Jacobian of the transformation:

$$f_Y(y) = f_X(g^{-1}(y))g'^{-1}(y) \tag{9.5}$$

$$= f_X(g^{-1}(y))|\frac{dx}{dy}| \tag{9.6}$$

In this case, $\phi = \theta^2$:

$$f_\Phi(\phi) = f_\theta(\sqrt{\phi})\frac{1}{2}\phi^{-\frac{1}{2}} \tag{9.7}$$

$$= 1 \times \frac{1}{2}\phi^{-\frac{1}{2}} \tag{9.8}$$

$$= \frac{1}{2}\phi^{-\frac{1}{2}} \tag{9.9}$$

Alternatively do this by sampling from a uniform prior for θ in R, then squaring each result,

```
fThetaSquared <- function(aNumSamples){
  lThetaSquared <- vector(length=aNumSamples)
  for(i in 1:aNumSamples){
    theta <- rbeta(1, 1, 1)
    lThetaSquared[i] <- theta ^ 2
  }
  return(lThetaSquared)
}
# Draw samples and graph result
theta <- fThetaSquared(100000)
hist(theta, 100, xlab="theta-squared")
```

9.2 Epilepsy

In the data file `conjugate_epil.csv` there is a count of seizures for 112 patients with epilepsy who took part in a study [2]. Assume a) the underlying rate of seizures is the same across all patients, and b) the event of a seizure occurring is independent of any other seizures occurring.

Problem 9.2.1. Under these assumptions what model might be appropriate for this data?

A Poisson distribution.

Problem 9.2.2. Write down the likelihood for the data.

The likelihood for a single observation x is given by:

$$L(\theta|x) = \frac{\theta^x e^{-\theta}}{x!} \tag{9.10}$$

For a data vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ if we assume independence between our observations we have:

$$L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} \tag{9.11}$$

Problem 9.2.3. Show that a gamma prior is conjugate to this likelihood.

The gamma distribution has the functional form:

$$p(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta} \tag{9.12}$$

The posterior then has the functional form:

$$\begin{aligned} p(\theta|\boldsymbol{x}) &\propto \theta^{\alpha-1} e^{-\beta\theta} \times \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} \\ &\propto \theta^{\alpha-1+\sum_{i=1}^{n} x_i} \times e^{-(\beta+n)\theta} \end{aligned}$$

Which is the same θ dependence as a $\Gamma(\alpha+\sum_{i=1}^{n} x_i, \beta+n)$ distribution $\implies$ this must be the posterior distribution! Therefore the posterior is a gamma distribution as well as the prior $\therefore$ conjugate.

Problem 9.2.4. Assuming a $\Gamma(4, 0.25)$ (with a parameterisation such that it has mean of 16) prior. Find the posterior distribution, and graph it.

See above problem for derivation of the posterior density. The graph of the posterior should look like Figure 9.6.

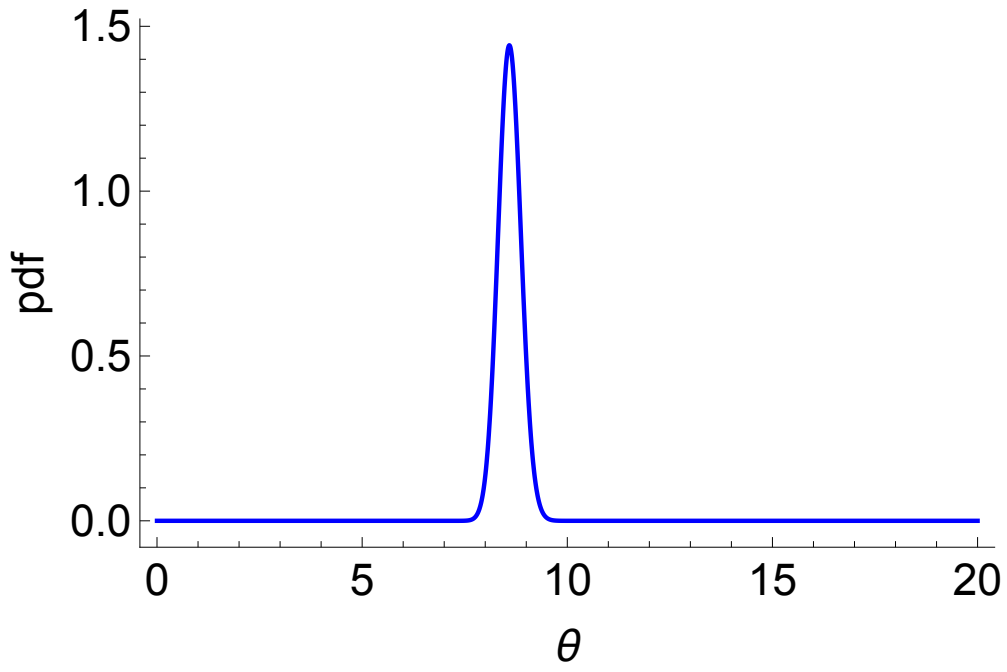

Figure 9.6: The posterior for the epilepsy example.

Problem 9.2.5. Find or look-up the posterior predictive distribution, and graph it.

The posterior predictive distribution is a negative binomial - this can be derived by:

$$\begin{aligned} p(\tilde{x}|x) &= \int p(\tilde{x}|\theta, \boldsymbol{x}) \times p(\theta|\boldsymbol{x}) \mathrm{d}\theta \\ &= \int p(\tilde{x}|\theta) \times p(\theta|\boldsymbol{x}) \mathrm{d}\theta \\ &\ldots \end{aligned}$$

where ... can be found via Googling. The posterior predictive distribution turns out to be $NB(\sum_{i=1}^{n} x_i + \alpha, \beta + n)$, where (α, β) are the parameters of the gamma prior distribution. The graph is shown in Figure 9.7.

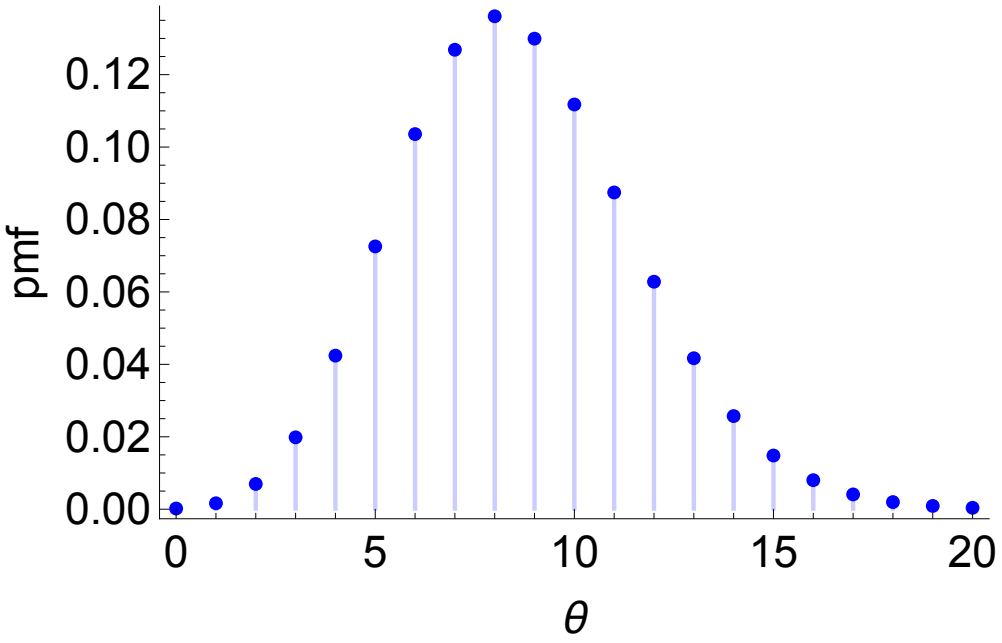

Figure 9.7: Posterior predictive distribution for seizure data.

Problem 9.2.6. Comment on the suitability of the model to the data.

In Figure 9.8 we see that the real data is much more dispersed than the simulated. This is likely for a number of reasons: for example, the event of a seizure is not likely independent of others (they come in clusters); also the rate of seizures varies between subjects (in other words the data are not exchangeable). Amongst other reasons these suggest that a Poisson model is not well suited here, and we would be better off using a more robust distribution for the likelihood, for example the negative binomial.

Figure 9.8: Comparing actual vs simulated seizures.

9.3 Light speed

The data file `conjugate_newcomb.csv` provides Simon Newcombs (1882) measurements of the passage time (in millionths of a second) it took light to travel from his lab to a mirror on the Washington Monument, and back again. The distance of the path travelled is about 7.4km. The primary goal of this experiment is to determine the speed of light, and to quantify the uncertainty of the measurement. We assume there are a multitude of factors that additively result in measurement error for the passage time.

Problem 9.3.1. Why might a normal distribution be appropriate here?

There are a range of factors that influence the measurement of the passage time. If these factors are roughly independent, and they affect the measurement additively, then the (Lindberg-Lévy) central limit theorem applies.

Problem 9.3.2. Write down the likelihood for all the data.

The likelihood of a single data point x is given by:

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{9.13}$$

If we assume measurements are independent, and identically-distributed then we just need to multiply together the individual likelihoods:

$$L(\mu, \sigma|\boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{9.14}$$

Problem 9.3.3. Derive the maximum likelihood estimators of all parameters.

It's easiest to first take the log:

$$l(\mu, \sigma|\boldsymbol{x}) = -\frac{n}{2} log(2\pi) - \frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{9.15}$$

Then maximising this function over (μ, σ^2), we find that:

$$\hat{\mu} = \bar{x}$$
$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Problem 9.3.4. Based on the likelihood function what functional form for the prior $p(\mu, \sigma^2)$ would make it conjugate?

We want a prior that when multiplied by a normal gives a distribution of the same family. There are a few choices here, but the only one that is a valid probability distribution is a normal-inverse-gamma or normal-inverse-chi-squared (they are both the same thing).

Problem 9.3.5. Assuming a decomposition of the prior $p(\mu, \sigma^2) = p(\sigma^2) \times p(\mu|\sigma^2)$, what priors might we use?

Again a normal inverse gamma. You could use an improper $p(\sigma^2) \propto \frac{1}{\sigma^2}$ but it's better to use fully-valid probability distributions.

Problem 9.3.6. (Difficult) Using these priors, find the parameters of the posterior distribution.

Look it up in Gelman [1].

Problem 9.3.7. Comment on the suitability of the model to the data. (You can use the ML estimates here, or if you're feeling ambitious, the full posterior predictive distribution.)

Using the posterior predictive simulate data and compare with the actual we see that the normal distribution is not sufficiently robust. We would be better using a Student t distribution.

Bibliography

[1] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[2] Peter F Thall and Stephen C Vail. Some covariance models for longitudinal count data with overdispersion. *Biometrics*, pages 657–671, 1990.