Chapter 10

Evaluation of model fit and hypothesis testing

10.1 WHO's reported novel disease outbreaks

Suppose that you are interested in modelling the number of outbreaks of novel diseases that the WHO reports each year. Since these outbreaks are of new diseases, you assume that you can model the outbreaks as **independent** events, and hence decide to use a Poisson likelihood; $X_t \sim \text{Poisson}(\lambda)$, where X_t is the number of outbreaks in year t, and λ is the mean number of outbreaks.

Problem 10.1.1. You decide to use a $\Gamma(3, 0.5)$ prior for the mean parameter (λ) of your Poisson likelihood (where a $\Gamma(\alpha, \beta)$ is defined to have a mean of $\frac{\alpha}{\beta}$). Graph this prior.

This can be done in R using the following command,

```
curve(dgamma(x, 3, 0.5), 0, 20, xlab='lambda', ylab='pdf')
```

Problem 10.1.2. Suppose that the number of new outbreaks over the past 5 years is $X = (3, 7, 4, 10, 11)$. Using the conjugate prior rules for a Poisson distribution with a gamma prior, find the posterior and graph it.

Hint: look at Table 9.1 in the main text.

The posterior distribution is given by a $\Gamma(3 + \sum_{t=1}^{5} X_t, 0.5 + 5)$ distribution. This can be graphed in R using,

```
X <- c(3, 7, 4, 10, 11)
curve(dgamma(x, 3 + sum(X), 0.5 + length(X)),0 , 20, xlab='lambda', ylab='pdf')
```

It has a peak at $<\sim 7$, near to the mean of the data.

Problem 10.1.3. Generate 10,000 samples from the posterior predictive distribution, and graph the distribution. To do this we first independently sample a value λ_i from the posterior distribution, then sample a value of X from a $Poisson(\lambda_i)$ distribution. We carry out this process 10,000 times.

Hint: use R's `rgamma` and `rpois` functions to draw (pseudo-)independent samples from the gamma and Poisson distributions respectively.

I prefer to do this by creating a function in R that implements the above then plots the result,

```
fPosteriorPredictive <- function(numSamples, alpha, beta){
  X <- vector(length=numSamples)
  for(i in 1:numSamples){
    aLambda <- rgamma(1, alpha, beta)
    X[i] <-rpois(1, aLambda)
  }
  return(X)
}

PPC.X <- fPosteriorPredictive(10000, 3 + sum(X), 0.5 + length(X))
hist(PPC.X, xlab='X', main='10,000 posterior predictive samples')
```

Problem 10.1.4. Compare the actual data with your 10,000 posterior predictive samples. Does your model fit the data?

The most extreme points of the data are the years with 3 and 11 outbreaks respectively. We can compare our posterior predictive samples with these extrema in R,

```
mean(PPC.X >= 11)
mean(PPC.X <= 3)
```

and find that roughly 10% of samples are greater than or equal to 11, and approximately the same proportion are less than or equal to 3. These Bayesian p values aren't too close to 0, and so our data appears to fit the data reasonably well.

Problem 10.1.5. (Optional) Can you think of a better posterior predictive check to carry out on the data?

A better posterior predictive check would generate 10,000 samples of 5 observations, and count the number where the minimum point is 3 **and** the maximum is 11 (or more extreme). To do this I implemented a new function,

```
fPosteriorPredictiveGeneral <- function(numObsPerSample, numSamples,
                                        alpha, beta){
  X <- matrix(nrow=numSamples, ncol=numObsPerSample)
  for(i in 1:numSamples){
    aLambda <- rgamma(1, alpha, beta)
    X[i, ] <-rpois(numObsPerSample, aLambda)
  }
  return(X)
}

aNumSamples <- 10000
```

```
PPC.better <- fPosteriorPredictiveGeneral(5, aNumSamples,
                                          3 + sum(X), 0.5 + length(X))
lIndicator <- vector(length=aNumSamples)
for(i in 1:aNumSamples)
  lIndicator[i] <- ifelse(min(PPC.better[i, ]) <= 3 &
                          max(PPC.better[i, ]) >= 11,
                          1, 0)
mean(lIndicator)
```

and you should get about 10% here. So it still looks like our model fits the data ok.

Problem 10.1.6. The WHO issues a press release where they state that the number of novel disease outbreaks for this year was 20. Use your posterior predictive samples to test whether your model is a good fit to the data.

Since we are just looking at a single data point we can use our simpler posterior predictive function to generate samples (or just reuse the previously-generated sample),

```
fPosteriorPredictive <- function(numSamples, alpha, beta){
  X <- vector(length=numSamples)
  for(i in 1:numSamples){
    aLambda <- rgamma(1, alpha, beta)
    X[i] <-rpois(1, aLambda)
  }
  return(X)
}

PPC.X <- fPosteriorPredictive(10000, 3 + sum(X), 0.5 + length(X))
mean(PPC.X >= 20)
```

where you should obtain a p value of less than 1%, indicating model misfit. This is a test of out-of-sample predictive capability, and so we would expect this p value to be more extreme than the within-sample one that we calculate below.

Problem 10.1.7. By using your previously determined posterior as a prior, update your posterior to reflect the new datum. Graph the PDF for this new distribution.

The new posterior here is a $\Gamma(3 + 35 + 20, 0.5 + 5 + 1)$ distribution,

```
curve(dgamma(x, 3 + sum(X) + 20, 0.5 + 5 + 1),
      0, 20, xlab='lambda', ylab='pdf')
```

Problem 10.1.8. Generate posterior predictive samples from your new posterior and use it to test the validity of your model.

Here I would generate 10,000 samples of 6 observations and count the number of times that you generate 20 or more cases in a particular year.

```
PPC.better <- fPosteriorPredictiveGeneral(6, aNumSamples,
                                          3 + sum(X) + 20,
                                          0.5 + 5 + 1)
lIndicator <- vector(length=aNumSamples)
for(i in 1:aNumSamples)
  lIndicator[i] <- ifelse(max(PPC.better[i, ]) >= 20, 1, 0)
mean(lIndicator)
```

where again the p value is less than 5% and hints at model misfit. This a within-sample measure of predictive capability of the model.

Problem 10.1.9. Would you feel comfortable using this model to predict the number of disease outbreaks next year?

No! Even the within-sample prediction is poor. It's probably that some of these outbreaks are related to one another – either they are different strains from a common disease, or they are the result of a common exogenous shock (e.g. civil war).

10.2 Sleep-deprived reactions

These data are from a study described in Belenky et al. (2003) [2] that measured the effect of sleep deprivation on cognitive performance. Eighteen subjects were chosen from a population of interest (lorry drivers) who were restricted to 3 hours of sleep during the trial. On each day of the experiment their reaction time to a visual stimulus was measured. The data for this example is contained in `evaluation_sleepstudy.csv` and consists of three variables, *Reaction*, *Days* and *Subject ID*, which measure the reaction time of a given subject on a particular day.

A simple model that explains the variation in reaction times is a linear regression model of the form:

$$R(t) \sim \mathcal{N}(\alpha + \beta t, \sigma) \tag{10.1}$$

where $R(t)$ is the reaction time on day t of the experiment across all observations.

Problem 10.2.1. By graphing all the data, critically assess the validity of the model to the data.

A simple graph of the time against reaction time is a first starter here. From this it looks like there may be some heteroscedasticity (higher variance) at later times. This can be done in R using,

```
library(ggplot2)
df <- read.csv('evaluation_sleepstudy.csv')
ggplot(data=df, aes(x=Days, y=Reaction)) + geom_point() +
  geom_smooth(method='lm')
```

Problem 10.2.2. Graph the data at the individual subject data using R's "lattice" package, or otherwise. What does this suggest about assuming a common β across all participants?

Using a lattice plot (see Figure 10.1),

```
xyplot(Reaction ~ Days | Subject, df, type=c("g", "p", "r"),
       index=function(x, y) coef(lm(y ~ x))[1],
       xlab="Days of sleep deprivation",
       ylab="Average reaction time (ms)",
       aspect="xy")
```

From an examination of the data at this level it is clear that there is considerable variability in the performance of the participants. As such, any attempts to lump the data together and apply a single analysis to it are going to suffer from considerable participant-level biases.

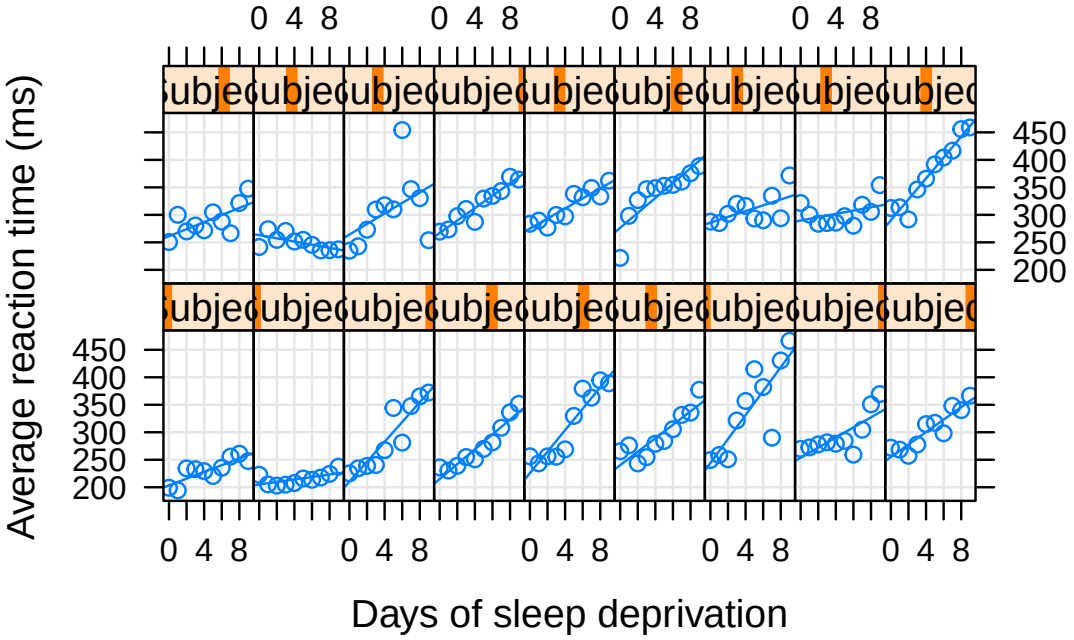

Figure 10.1: Reaction times versus days of sleep deprivation at the participant level.

Problem 10.2.3. The above model has been fit to the data using MCMC, with 2000 samples from the posterior distribution for (α, β, σ) contained in the file `evaluation_sleepPosteriors.csv`. Generate samples from the posterior predictive distribution, and visualise them in an appropriate way.

These are shown in Figure 10.2. It is important here to show the time aspect of the data; just lumping it all together in a histogram misses the point.

Problem 10.2.4. How does the posterior predictive data compare to the actual data?

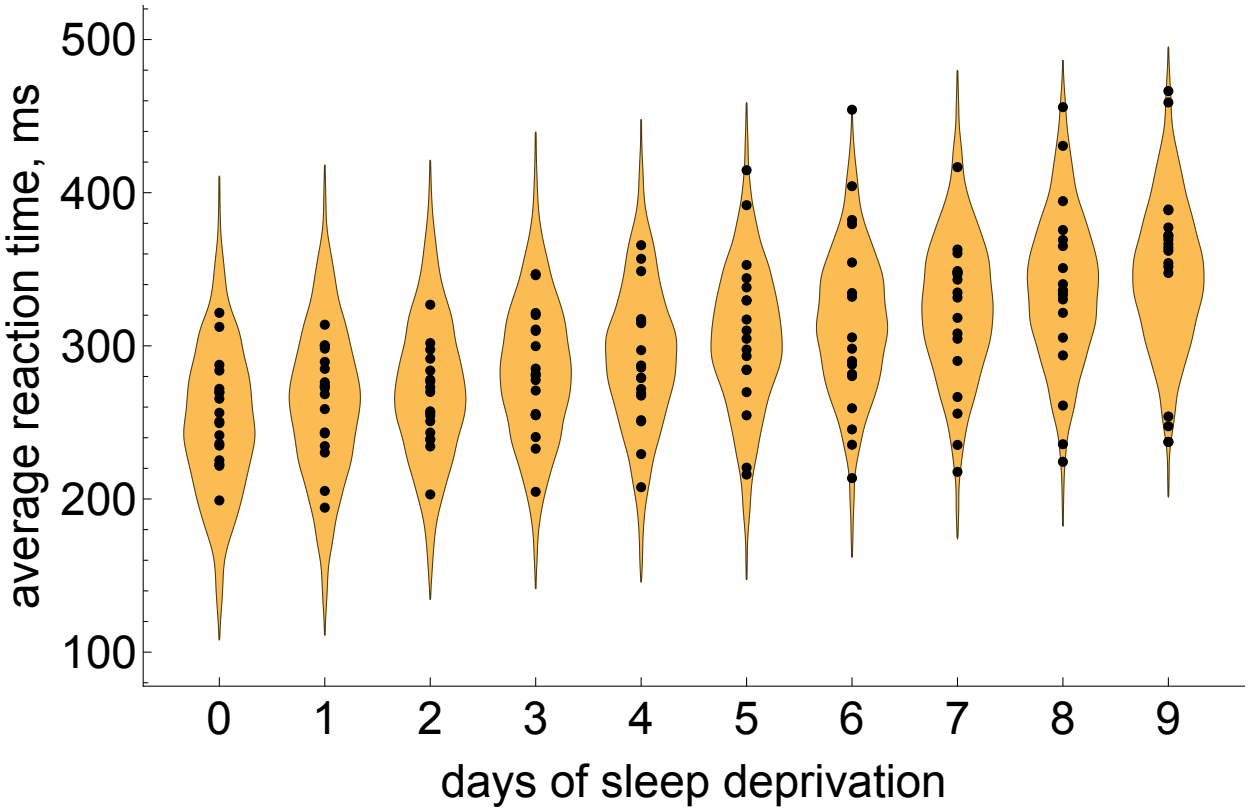

Figure 10.2: Posterior predictive distributions (orange) vs data (black).

The key here is to look at the data at the subject level. Averaging over all subjects makes it look like our model is doing ok, but this masks the (sometimes) very poor performance at the individual subject level (see Figure 10.3 for one example of this for subject 310).

Problem 10.2.5. How (if at all) do the posterior predictive checks suggest we need to change our model?

Hierarchical model where we allow there to be inter-subject variability in the effect of sleep deprivation on reaction time (β).

10.3 Discoveries data

The file `evaluation_discoveries.csv` contains data on the numbers of "great" inventions and scientific discoveries in each year from 1860 to 1959 [1]. The aim of this problem is for you to build a statistical model that provides a reasonable approximation to this series. As such, you will need to choose a likelihood, specify a prior on any parameters, and go through and calculate a posterior. Once you have a posterior, you will want to carry out posterior predictive checks to see that your model behaves as desired.

Answer: first plot the data! Both a time series and histogram are useful here (see Figure 10.4). To me the left hand plot suggests that there is some temporal autocorrelation in the data (perhaps

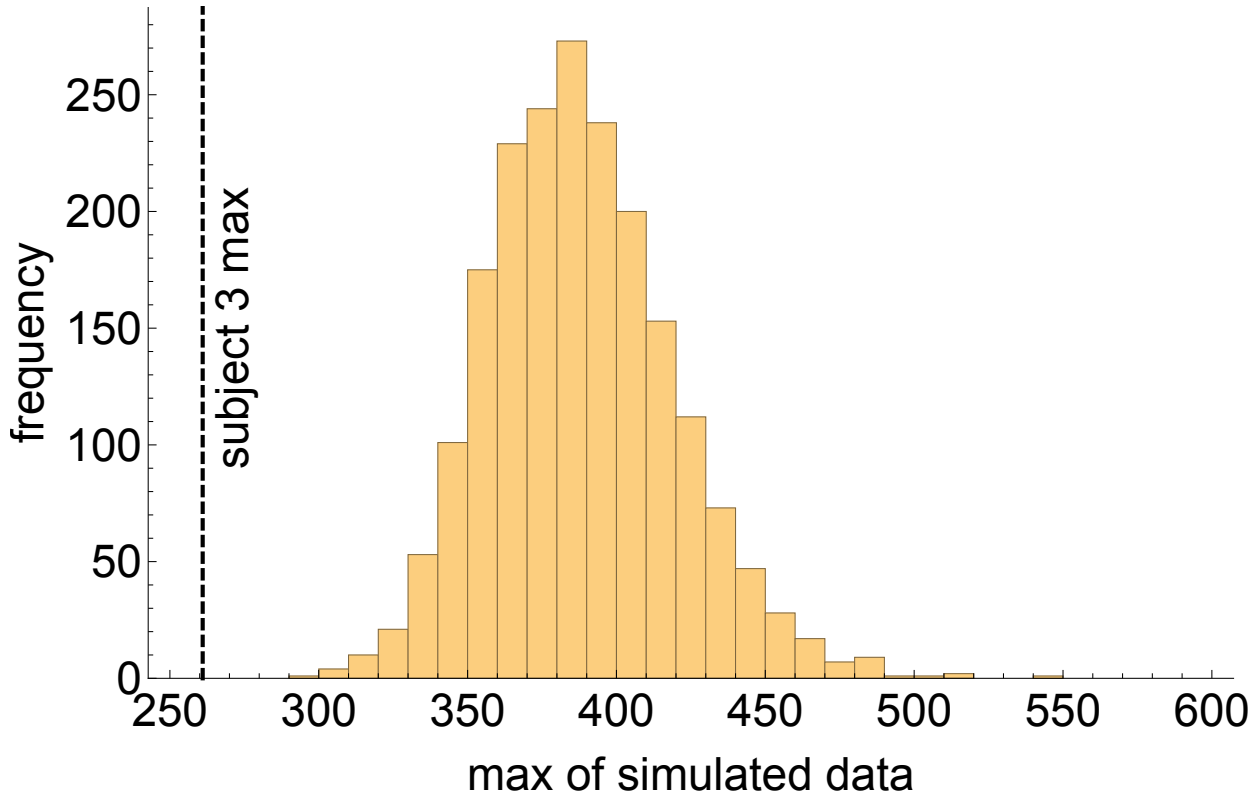

Figure 10.3: Posterior predictive simulated max (orange) versus the maximum of subject 310 (dashed line).

invalidating an assumption of independence, and/or identical distribution). The histogram would seem to support this claim, since the variance is fairly obviously greater than the mean. I also plot an autocorrelogram of the data which suggests that there is autocorrelation in the series.

Now make some assumptions about the occurrence of discoveries; namely that they are independent and identically-distributed over time. Both of these assumptions may be suspect: independence may be violated (as I indicate above) if one discovery leads to another; identical distribution may be invalidated if technological progress leads to an increased rate of discoveries at some points in time.

However, it is not a bad idea to start with making these assumptions, under the supposition that they may be suspect. Our aim is to make the simplest model that explains the data, and so we don't want to jump straight to a more complex model unless we know for sure that the simple one fails.

If we do make the above assumptions then a Poisson model is a reasonable starting point. If we use a Poisson model, then we may as well use the conjugate prior; a gamma distribution. The results of assuming this framework are shown in Figure 10.5; where we see a tight posterior centred around a mean of 3 discoveries per year.

Carrying out some PPCs here using the posterior predictive distribution from the Poisson likelihood model we find that our model is *unable* to generate the same amount of variation seen in the data (Figure 10.6). This suggests that one or more of the assumptions on which our data are based are

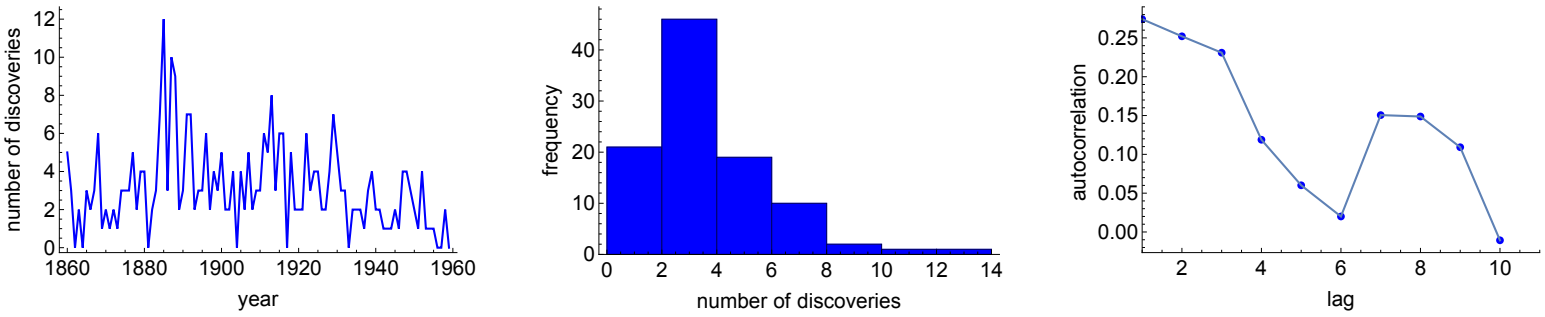

Figure 10.4: Characteristics of the "discoveries" data set.

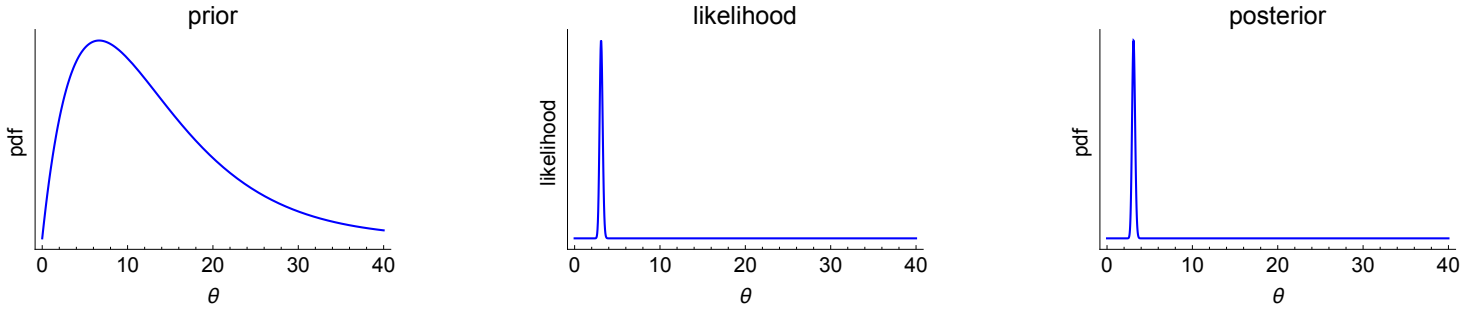

Figure 10.5: Prior, likelihood and posterior for a Poisson likelihood and $\Gamma(2, 0.15)$ prior for the discoveries dataset.

invalid.

There are multiple ways forward from here. To me there are two approaches that "jump out": **a.** use a sampling distribution that allows for non-independent events, but does not explicitly model the cycles of discovery; **b.** explicitly model the latent rate of discovery rate. Approach **a.** would suggests a negative binomial likelihood, and would certainly allow for the range in the data to be replicated well. However, I fear that such an approach - by ignoring the fact that the rate of discoveries changes through time - would fail to capture the intervals of high discovery rate that we see in the data. In other words those times (for example, between 1880 and 1890) where there is a persistently high rate of discovery. Approach **b.** would be more comprehensive and would perhaps use a negative binomial sampling model for each year, but allow its mean to vary over time. So if we imagine that the mean of the process at time t is θ_t, then we might assume:

$$\theta_t = \rho\theta_{t-1} + \epsilon_t \tag{10.2}$$

So an AR(1) process explicitly. both of these approaches favour a MCMC approach (particularly the AR(1) process one). As such, I have not tried either these investigations myself, as I wouldn't necessarily expect a student to be able to do these at this stage.

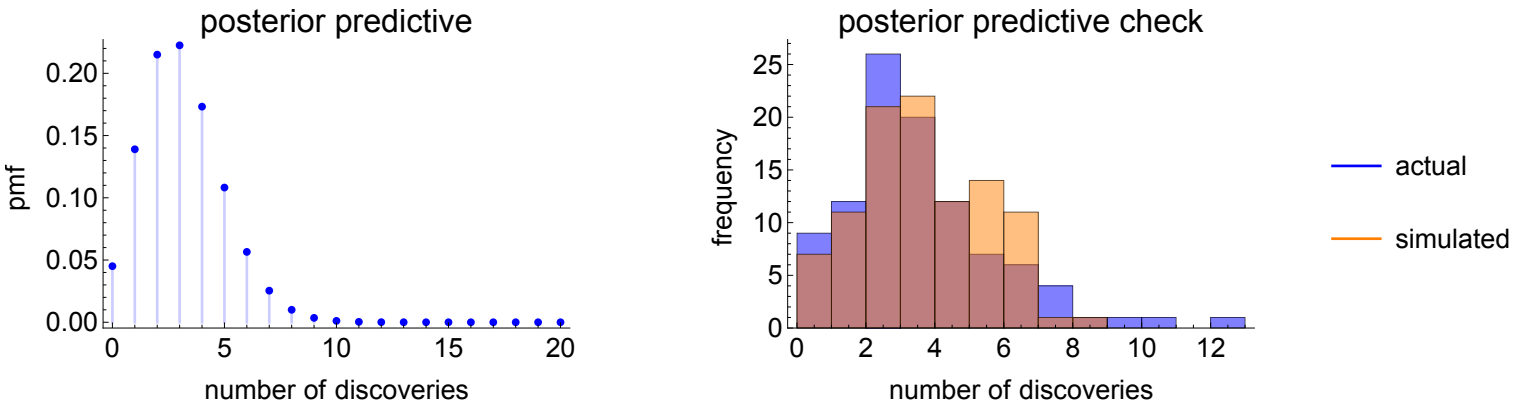

Figure 10.6: The posterior predictive distribution (left) and a posterior predictive comparison of the actual data with a simulated set (right).

10.4 Marginal likelihood of voting

Suppose that we collect survey data where respondents are asked to indicate for whom they will vote in an upcoming election. Each poll consists of a sample size of 10 and we collect the following data for 20 such polls: $\{2, 7, 4, 5, 4, 5, 6, 4, 4, 4, 5, 6, 5, 7, 6, 2, 4, 6, 6, 6\}$. We model each outcome as having been obtained from a $X_i \sim \mathcal{B}(10, \theta)$ distribution.

Problem 10.4.1. Find the posterior distribution where we specify $\theta \sim beta(a, 1)$ as a prior. Graph how the posterior changes as $a \in [1, 10]$.

The posterior distribution is given by (because of conjugacy): $\theta \sim beta(a + \sum X_i, 1 + \sum N_i - \sum X_i)$. The graph of the posterior as a function of a is shown in Figure 10.7, where we note the relative insensitivity to the prior.

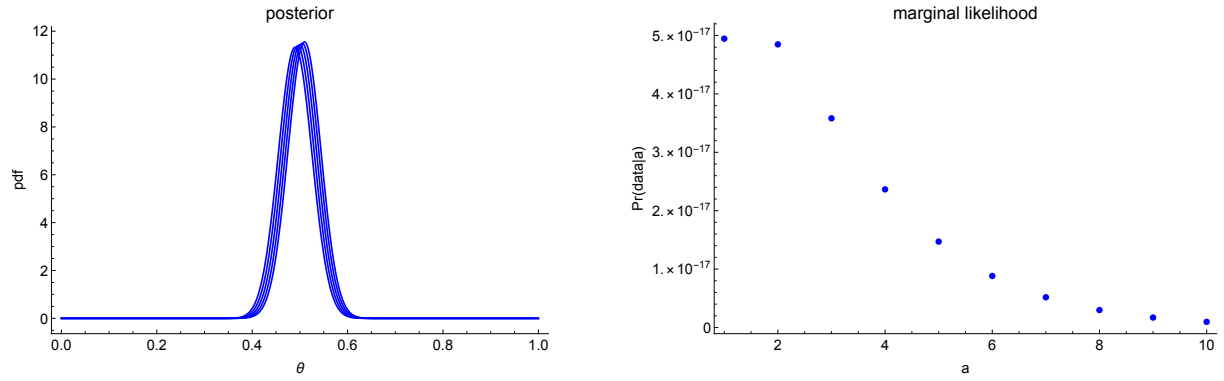

Figure 10.7: The posterior distribution (left) and the marginal likelihood (right) as a function of prior parameter a. Here the prior specified is $\theta \sim beta(a, 1)$. In the left hand graph the different lines correspond to different choices of a.

Problem 10.4.2. Graph the marginal likelihood as a is increased between 1 and 10 (just use integer values).

See Figure 10.7.

Problem 10.4.3. Calculate the Bayes factor where we compare the model where $a = 1$ to that when $a = 10$? Hence comment on the use of Bayes factors as a method for choosing between competing models.

This is approximately,

$$BF = \frac{4.94 \times 10^{-17}}{9.64 \times 10^{-19}} \approx 51. \tag{10.3}$$

So we see that there is a strong sensitivity of Bayes factors to choice of priors, even if the posterior is relatively insensitive.

Bibliography

[1] *The World Almanac and Book of Facts*. 1975.

[2] Gregory Belenky, Nancy J Wesensten, David R Thorne, Maria L Thomas, Helen C Sing, Daniel P Redmond, Michael B Russo, and Thomas J Balkin. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of sleep research*, 12(1):1–12, 2003.