4.3 A Multinomial Model

Gelman et al. (2003) describe a sample survey conducted by CBS News before the 1988 presidential election. A total of 1447 adults were polled to indicate their preference; $y_1 = 727$ supported George Bush, $y_2 = 583$ supported Michael Dukakis, and $y_3 = 137$ supported other candidates or expressed no opinion. The counts y_1, y_2, and y_3 are assumed to have a multinomial distribution with sample size n and respective probabilities θ_1, θ_2, and θ_3. If a uniform prior distribution is assigned to the multinomial vector $\theta = (\theta_1, \theta_2, \theta_3)$, then the posterior distribution of θ is proportional to

$$g(\theta) = \theta_1^{y_1} \theta_2^{y_2} \theta_3^{y_3},$$

which is recognized as a Dirichlet distribution with parameters $(y_1 + 1, y_2 + 1, y_3 + 1)$. The focus is to compare the proportions of voters for Bush and Dukakis by considering the difference $\theta_1 - \theta_2$.

The summarization of the Dirichlet posterior distribution is again conveniently done by simulation. Although the base R package does not have a function to simulate Dirichlet variates, it is easy to write a function to simulate this distribution based on the fact that if W_1, W_2, W_3 are independentlt distributed from gamma$(\alpha_1, 1)$, gamma$(\alpha_2, 1)$, gamma$(\alpha_3, 1)$ distributions and $T = W_1 + W_2 + W_3$, then the distribution of the proportions $(W_1/T, W_2/T, W_3/T)$ has a Dirichlet$(\alpha_1, \alpha_2, \alpha_3)$ distribution. The R function `rdirichlet` in the package `LearnBayes` uses this transformation of random variates to simulate draws of a Dirichlet distribution. One thousand vectors θ are simulated and stored in the matrix `theta`.

```
> alpha = c(728, 584, 138)
> theta = rdirichlet(1000, alpha)
```

Since we are interested in comparing the proportions for Bush and Dukakis, we focus on the difference $\theta_1 - \theta_2$. A histogram of the simulated draws of this difference is displayed in Figure 4.2. Note that all of the mass of this distribution is on positive values, indicating that there is strong evidence that the proportion of voters for Bush exceeds the proportion for Dukakis.

```
> hist(theta[, 1] - theta[, 2], main="")
```

In the United States presidential election, there are 50 states plus the District of Columbia, and each has an assigned number of electoral votes. The candidate receiving the largest number of votes in a particular state receives the corresponding number of electoral votes, and for a candidate to be elected, he or she must receive a majority of the total number (538) of electoral votes. In the 2008 election between Barack Obama and John McCain, suppose we wish to predict the total number of electoral votes EV_O obtained by Obama. Let θ_{Oj} and θ_{Mj} denote the proportion of voters respectively for Obama and

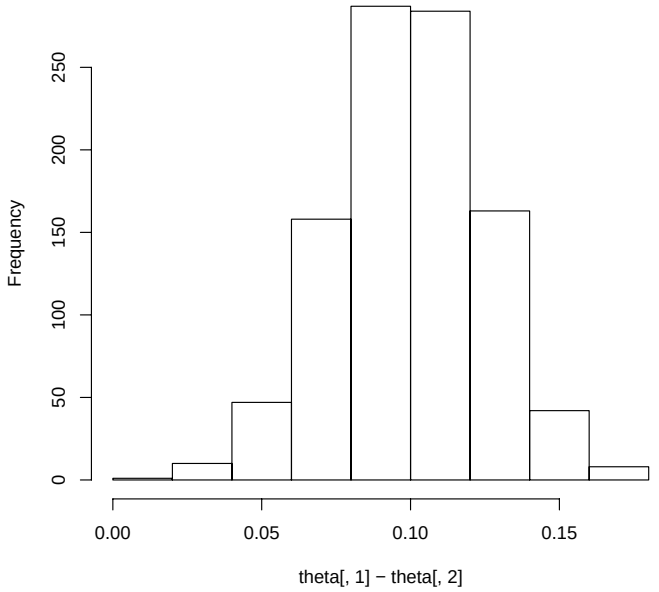

Fig. 4.2. Histogram of simulated sample of the marginal posterior distribution of $\theta_1 - \theta_2$ for the multinomial sampling example.

McCain in the jth state. One can express the number of electoral votes for Obama as

$$EV_O = \sum_{j=1}^{51} EV_j I(\theta_{Oj} > \theta_{Mj}),$$

where EV_j is the number of electoral votes in the jth state and $I()$ is the indicator function, which is equal to 1 if the argument is true and 0 otherwise.

On the Sunday before Election Day, the website www.cnn.com gives the results of the most recent poll in each state. Let q_{Oj} and q_{Mj} denote the sample proportions of voters for Obama and McCain in the ith state. We make the conservative assumption that each poll is based on a sample of 500 voters. Assuming a uniform prior on the vector of proportions, the vectors $(\theta_{O1}, \theta_{M1}), ..., (\theta_{O51}, \theta_{M51})$ have independent posterior distributions, where the proportions favoring the candidates in the ith state, $(\theta_{Oi}, \theta_{Mi}, 1 - \theta_{Oi}, \theta_{Mi})$, have a Dirichlet distribution with parameters $(500q_{Oj} + 1, 500q_{Mj} + 1, 500(1 - q_{Oj} - q_{Mj}) + 1)$.

Based on the posterior distribution of the state proportions, one can simulate from the posterior distribution of the electoral votes for Obama. The

dataset `election.2008` in the `LearnBayes` package contains for each state the percentage of voters in the poll for McCain `M.pct`, the percentage of voters in the poll for Obama `O.pct`, and the number of electoral votes `EV`.

```
> library(LearnBayes)
> data(election.2008)
> attach(data)
```

We write a short function `prob.Obama` that will use simulation from the Dirichlet distributions to compute the posterior probability that θ_{Oj} exceeds θ_{Mj} in the jth state.

```
> prob.Obama=function(j)
+ {
+ p=rdirichlet(5000,
+   500*c(M.pct[j],O.pct[j],100-M.pct[j]-O.pct[j])/100+1)
+ mean(p[,2]>p[,1])
+ }
```

We compute this Obama win probability for all states by using the `sapply` function.

```
> Obama.win.probs=sapply(1:51,prob.Obama)
```

Now that we have the win probabilities, we can simulate from the posterior distribution of the Obama electoral votes by flipping a set of 51 biased coins, where the coin probabilities correspond to the Obama state win probabilities. Then we compute the number of Obama electoral votes based on the results of the coin flips. We implement one simulation using the function `sim.election` and repeat this simulation 1000 times using the `replicate` function. The vector `sim.EV` contains the number of electoral votes in the simulations.

```
> sim.election=function()
+ {
+ winner=rbinom(51,1,Obama.win.probs)
+ sum(EV*winner)
+ }
> sim.EV=replicate(1000,sim.election())
```

We construct a histogram of the posterior of EV_O, which is displayed in Figure 4.3.

```
> hist(sim.EV,min(sim.EV):max(sim.EV),col="blue")
> abline(v=365,lwd=3)  # Obama received 365 votes
> text(375,30,"Actual \n Obama \n total")
```

The actual Obama electoral vote total of 365 is displayed on the graph. It would have been possible to improve our prediction by using more data than just the results of a single poll in each state. But the actual electoral vote total did fall within the 90% equal-tail prediction interval.

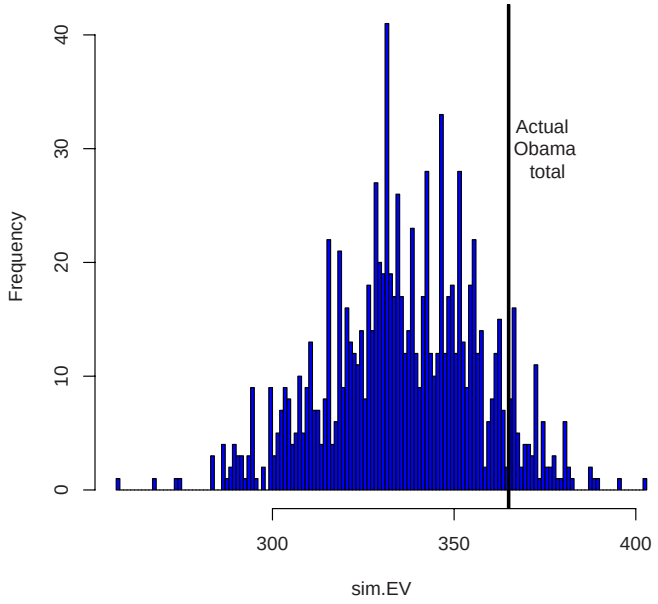

Fig. 4.3. Histogram of 1000 simulated draws of the total electoral vote for Barack Obama in the 2008 U.S. presidential election. The actual electoral vote of 365 is indicated by a vertical line.

4.4 A Bioassay Experiment

In the development of drugs, bioassay experiments are often performed on animals. In a typical experiment, various dose levels of a compound are administered to batches of animals and a binary outcome (positive or negative) is recorded for each animal. We consider data from Gelman et al. (2003), where one observes a dose level (in log g/ml), the number of animals, and the number of deaths for each of four groups. The data are displayed in Table 4.1.

Table 4.1. Data from the bioassay experiment.

Dose	Deaths	Sample Size
−0.86	0	5
−0.30	1	5
−0.05	3	5
0.73	5	5