Sampling from the joint posterior distribution

To sample from the joint posterior distribution, just as in the previous section, we first draw σ^2 from its marginal posterior distribution (3.9), then draw μ from its normal conditional posterior distribution (3.8), using the simulated value of σ^2.

Analytic form of the marginal posterior distribution of μ

Integration of the joint posterior density with respect to σ^2, in a precisely analogous way to that used in the previous section, shows that the marginal posterior density for μ is

$$\begin{aligned} p(\mu|y) &\propto \left(1+\frac{\kappa_n(\mu-\mu_n)^2}{\nu_n\sigma_n^2}\right)^{-(\nu_n+1)/2} \\ &= t_{\nu_n}(\mu|\mu_n,\sigma_n^2/\kappa_n). \end{aligned}$$

3.4 Multinomial model for categorical data

The binomial distribution that was emphasized in Chapter 2 can be generalized to allow more than two possible outcomes. The multinomial sampling distribution is used to describe data for which each observation is one of k possible outcomes. If y is the vector of counts of the number of observations of each outcome, then

$$p(y|\theta) \propto \prod_{j=1}^{k} \theta_j^{y_j},$$

where the sum of the probabilities, $\sum_{j=1}^k \theta_j$, is 1. The distribution is typically thought of as implicitly conditioning on the number of observations, $\sum_{j=1}^k y_j = n$. The conjugate prior distribution is a multivariate generalization of the beta distribution known as the Dirichlet,

$$p(\theta|\alpha) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j-1},$$

where the distribution is restricted to nonnegative θ_j's with $\sum_{j=1}^k \theta_j = 1$; see Appendix A for details. The resulting posterior distribution for the θ_j's is Dirichlet with parameters $\alpha_j + y_j$.

The prior distribution is mathematically equivalent to a likelihood resulting from $\sum_{j=1}^k \alpha_j$ observations with α_j observations of the jth outcome category. As in the binomial there are several plausible noninformative Dirichlet prior distributions. A uniform density is obtained by setting $\alpha_j = 1$ for all j; this distribution assigns equal density to any vector θ satisfying $\sum_{j=1}^k \theta_j = 1$. Setting $\alpha_j = 0$ for all j results in an improper prior distribution that is uniform in the $\log(\theta_j)$'s. The resulting posterior distribution is proper if there is at least one observation in each of the k categories, so that each component of y is positive. The bibliographic note at the end of this chapter points to other suggested noninformative prior distributions for the multinomial model.

Example. Pre-election polling

For a simple example of a multinomial model, we consider a sample survey question with three possible responses. In late October, 1988, a survey was conducted by CBS News of 1447 adults in the United States to find out their preferences in the upcoming

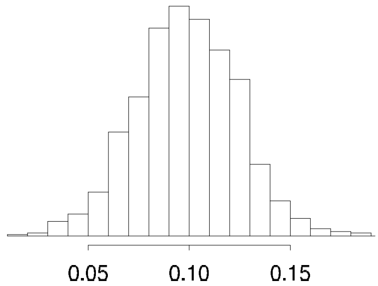

Figure 3.2 *Histogram of values of $(\theta_1 - \theta_2)$ for 1000 simulations from the posterior distribution for the election polling example.*

presidential election. Out of 1447 persons, $y_1 = 727$ supported George Bush, $y_2 = 583$ supported Michael Dukakis, and $y_3 = 137$ supported other candidates or expressed no opinion. Assuming no other information on the respondents, the 1447 observations are exchangeable. If we also assume simple random sampling (that is, 1447 names 'drawn out of a hat'), then the data (y_1, y_2, y_3) follow a multinomial distribution, with parameters $(\theta_1, \theta_2, \theta_3)$, the proportions of Bush supporters, Dukakis supporters, and those with no opinion in the survey population. An estimand of interest is $\theta_1 - \theta_2$, the population difference in support for the two major candidates.

With a noninformative uniform prior distribution on θ, $\alpha_1 = \alpha_2 = \alpha_3 = 1$, the posterior distribution for $(\theta_1, \theta_2, \theta_3)$ is Dirichlet$(728, 584, 138)$. We could compute the posterior distribution of $\theta_1 - \theta_2$ by integration, but it is simpler just to draw 1000 points $(\theta_1, \theta_2, \theta_3)$ from the posterior Dirichlet distribution and then compute $\theta_1 - \theta_2$ for each. The result is displayed in Figure 3.2. All of the 1000 simulations had $\theta_1 > \theta_2$; thus, the estimated posterior probability that Bush had more support than Dukakis in the survey population is over 99.9%.

In fact, the CBS survey does not use independent random sampling but rather uses a variant of a stratified sampling plan. We discuss an improved analysis of this survey, using some knowledge of the sampling scheme, in Section 8.3 (see Table 8.2 on page 207).

In complicated problems—for example, analyzing the results of many survey questions simultaneously—the number of multinomial categories, and thus parameters, becomes so large that it is hard to usefully analyze a dataset of moderate size without additional structure in the model. Formally, additional information can enter the analysis through the prior distribution or the sampling model. An informative prior distribution might be used to improve inference in complicated problems, using the ideas of hierarchical modeling introduced in Chapter 5. Alternatively, loglinear models can be used to impose structure on multinomial parameters that result from cross-classifying several survey questions; Section 16.7 provides details and an example.

3.5 Multivariate normal model with known variance

Here we give a somewhat formal account of the distributional results of Bayesian inference for the parameters of a multivariate normal distribution. In many ways, these results parallel those already given for the univariate normal model, but there are some important new aspects that play a major role in the analysis of linear models, which is the central activity of much applied statistical work (see Chapters 5, 14, and 15). This section can be viewed at this point as reference material for future chapters.