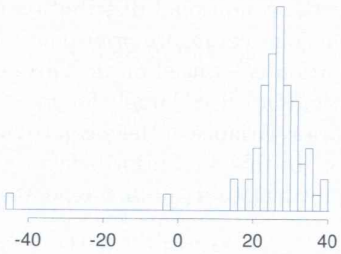

Figure 3.1 *Histogram of Simon Newcomb's measurements for estimating the speed of light, from Stigler (1977). The data are recorded as deviations from* 24,800 *nanoseconds.*

μ and variance σ^2. The main substantive goal is posterior inference for μ. The outlying measurements do not fit the normal model; we discuss Bayesian methods for measuring the lack of fit for these data in Section 6.3. The mean of the 66 measurements is $\overline{y} = 26.2$, and the sample standard deviation is $s = 10.8$. Assuming the noninformative prior distribution $p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$, a 95% central posterior interval for μ is obtained from the t_{65} marginal posterior distribution of μ as $\overline{y} \pm 1.997 s/\sqrt{66} = [23.6, 28.8]$.
The posterior interval can also be obtained by simulation. Following the factorization of the posterior distribution given by (3.5) and (3.3), we first draw a random value of $\sigma^2 \sim \text{Inv-}\chi^2(65, s^2)$ as $65s^2$ divided by a random draw from the χ^2_{65} distribution (see Appendix A). Then given this value of σ^2, we draw μ from its conditional posterior distribution, $\text{N}(26.2, \sigma^2/66)$. Based on 1000 simulated values of (μ, σ^2), we estimate the posterior median of μ to be 26.2 and a 95% central posterior interval for μ to be $[23.6, 28.9]$, close to the analytically calculated interval.
Incidentally, based on the currently accepted value of the speed of light, the 'true value' for μ in Newcomb's experiment is 33.0, which falls outside our 95% interval. This reinforces the fact that posterior inferences are only as good as the model and the experiment that produced the data.

3.3 Normal data with a conjugate prior distribution

A family of conjugate prior distributions

A first step toward a more general model is to assume a conjugate prior distribution for the two-parameter univariate normal sampling model in place of the noninformative prior distribution just considered. The form of the likelihood displayed in (3.2) and the subsequent discussion shows that the conjugate prior density must also have the product form $p(\sigma^2)p(\mu|\sigma^2)$, where the marginal distribution of σ^2 is scaled inverse-χ^2 and the conditional distribution of μ given σ^2 is normal (so that marginally μ has a t distribution). A convenient parameterization is given by the following specification:

$$\begin{aligned} \mu|\sigma^2 &\sim \text{N}(\mu_0, \sigma^2/\kappa_0) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2), \end{aligned}$$

which corresponds to the joint prior density

$$p(\mu, \sigma^2) \propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2]\right). \tag{3.6}$$

We label this the N-Inv-$\chi^2(\mu_0, \sigma_0^2/\kappa_0; \nu_0, \sigma_0^2)$ density; its four parameters can be identified as the location and scale of μ and the degrees of freedom and scale of σ^2, respectively.

The appearance of σ^2 in the conditional distribution of $\mu|\sigma^2$ means that μ and σ^2 are necessarily dependent in their joint conjugate prior density: for example, if σ^2 is large, then a high-variance prior distribution is induced on μ. This dependence is notable, considering that conjugate prior distributions are used largely for convenience. Upon reflection, however, it often makes sense for the prior variance of the mean to be tied to σ^2, which is the sampling variance of the observation y. In this way, prior belief about μ is calibrated by the scale of measurement of y and is equivalent to κ_0 prior measurements on this scale.

The joint posterior distribution, $p(\mu, \sigma^2|y)$

Multiplying the prior density (3.6) by the normal likelihood yields the posterior density

$$\begin{aligned} p(\mu, \sigma^2|y) &\propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu-\mu_0)^2]\right) \times \\ &\quad \times (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\overline{y}-\mu)^2]\right) \\ &= \text{N-Inv-}\chi^2(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2), \end{aligned} \tag{3.7}$$

where, after some algebra (see Exercise 3.9), it can be shown that

$$\begin{aligned} \mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\overline{y} \\ \kappa_n &= \kappa_0 + n \\ \nu_n &= \nu_0 + n \\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\overline{y}-\mu_0)^2. \end{aligned}$$

The parameters of the posterior distribution combine the prior information and the information contained in the data. For example μ_n is a weighted average of the prior mean and the sample mean, with weights determined by the relative precision of the two pieces of information. The posterior degrees of freedom, ν_n, is the prior degrees of freedom plus the sample size. The posterior sum of squares, $\nu_n\sigma_n^2$, combines the prior sum of squares, the sample sum of squares, and the additional uncertainty conveyed by the difference between the sample mean and the prior mean.

The conditional posterior distribution, $p(\mu|\sigma^2, y)$

The conditional posterior density of μ, given σ^2, is proportional to the joint posterior density (3.7) with σ^2 held constant,

$$\begin{aligned} \mu|\sigma^2, y &\sim \text{N}(\mu_n, \sigma^2/\kappa_n) \\ &= \text{N}\left(\frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\overline{y}}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}}\right), \end{aligned} \tag{3.8}$$

which agrees, as it must, with the analysis in Section 2.5 of μ with σ considered fixed.

The marginal posterior distribution, $p(\sigma^2|y)$

The marginal posterior density of σ^2, from (3.7), is scaled inverse-χ^2:

$$\sigma^2|y \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2). \tag{3.9}$$

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