This definition is formally vague since if we choose $\mathcal{P}$ as the class of all distributions, then $\mathcal{P}$ is always conjugate no matter what class of sampling distributions is used. We are most interested in *natural* conjugate prior families, which arise by taking $\mathcal{P}$ to be the set of all densities having the same functional form as the likelihood.

Conjugate prior distributions have the practical advantage, in addition to computational convenience, of being interpretable as additional data, as we have seen for the binomial example and will also see for the normal and other standard models in Sections 2.5 and 2.6.

Nonconjugate prior distributions

The basic justification for the use of conjugate prior distributions is similar to that for using standard models (such as binomial and normal) for the likelihood: it is easy to understand the results, which can often be put in analytic form, they are often a good approximation, and they simplify computations. Also, they will be useful later as building blocks for more complicated models, including in many dimensions, where conjugacy is typically impossible. For these reasons, conjugate models can be good starting points; for example, mixtures of conjugate families can sometimes be useful when simple conjugate distributions are not reasonable (see Exercise 2.4).

Although they can make interpretations of posterior inferences less transparent and computation more difficult, nonconjugate prior distributions do not pose any new conceptual problems. In practice, for complicated models, conjugate prior distributions may not even be possible. Section 2.4 and Exercises 2.10 and 2.11 present examples of nonconjugate computation; a more extensive nonconjugate example, an analysis of a bioassay experiment, appears in Section 3.7.

Conjugate prior distributions, exponential families, and sufficient statistics

We close this section by relating conjugate families of distributions to the classical concepts of exponential families and sufficient statistics. Readers who are unfamiliar with these concepts can skip ahead to the example with no loss.

Probability distributions that belong to an *exponential family* have natural conjugate prior distributions, so we digress at this point to review the definition of exponential families; for complete generality in this section, we allow data points y_i and parameters θ to be multidimensional. The class $\mathcal{F}$ is an exponential family if all its members have the form,

$$p(y_i|\theta) = f(y_i)g(\theta)e^{\phi(\theta)^T u(y_i)}.$$

The factors $\phi(\theta)$ and $u(y_i)$ are, in general, vectors of equal dimension to that of θ. The vector $\phi(\theta)$ is called the 'natural parameter' of the family $\mathcal{F}$. The likelihood corresponding to a sequence $y = (y_1, \ldots, y_n)$ of independent and identically distributed observations is

$$p(y|\theta) = \left(\prod_{i=1}^{n} f(y_i)\right) g(\theta)^n \exp\left(\phi(\theta)^T \sum_{i=1}^{n} u(y_i)\right).$$

For all n and y, this has a fixed form (as a function of θ):

$$p(y|\theta) \propto g(\theta)^n e^{\phi(\theta)^T t(y)}, \quad \text{where } t(y) = \sum_{i=1}^{n} u(y_i).$$

The quantity $t(y)$ is said to be a *sufficient statistic* for θ, because the likelihood for θ depends on the data y only through the value of $t(y)$. Sufficient statistics are useful in

algebraic manipulations of likelihoods and posterior distributions. If the prior density is specified as

$$p(\theta) \propto g(\theta)^{\eta} e^{\phi(\theta)^T \nu},$$

then the posterior density is

$$p(\theta|y) \propto g(\theta)^{\eta+n} e^{\phi(\theta)^T (\nu + t(y))},$$

which shows that this choice of prior density is conjugate. It has been shown that, in general, the exponential families are the only classes of distributions that have natural conjugate prior distributions, since, apart from certain irregular cases, the only distributions having a fixed number of sufficient statistics for all n are of the exponential type. We have already discussed the binomial distribution, where for the likelihood $p(y|\theta, n) = \text{Bin}(y|n, \theta)$ with n known, the conjugate prior distributions on θ are beta distributions. It is left as an exercise to show that the binomial is an exponential family with natural parameter $\text{logit}(\theta)$.

Example. Probability of a girl birth given placenta previa

As a specific example of a factor that may influence the sex ratio, we consider the maternal condition *placenta previa*, an unusual condition of pregnancy in which the placenta is implanted low in the uterus, obstructing the fetus from a normal vaginal delivery. An early study concerning the sex of placenta previa births in Germany found that of a total of 980 births, 437 were female. How much evidence does this provide for the claim that the proportion of female births in the population of placenta previa births is less than 0.485, the proportion of female births in the general population?

Analysis using a uniform prior distribution. Under a uniform prior distribution for the probability of a girl birth, the posterior distribution is Beta(438, 544). Exact summaries of the posterior distribution can be obtained from the properties of the beta distribution (Appendix A): the posterior mean of θ is 0.446 and the posterior standard deviation is 0.016. Exact posterior quantiles can be obtained using numerical integration of the beta density, which in practice we perform by a computer function call; the median is 0.446 and the central 95% posterior interval is $[0.415, 0.477]$. This 95% posterior interval matches, to three decimal places, the interval that would be obtained by using a normal approximation with the calculated posterior mean and standard deviation. Further discussion of the approximate normality of the posterior distribution is given in Chapter 4.

In many situations it is not feasible to perform calculations on the posterior density function directly. In such cases it can be particularly useful to use simulation from the posterior distribution to obtain inferences. The first histogram in Figure 2.3 shows the distribution of 1000 draws from the Beta(438, 544) posterior distribution. An estimate of the 95% posterior interval, obtained by taking the 25th and 976th of the 1000 ordered draws, is $[0.415, 0.476]$, and the median of the 1000 draws from the posterior distribution is 0.446. The sample mean and standard deviation of the 1000 draws are 0.445 and 0.016, almost identical to the exact results. A normal approximation to the 95% posterior interval is $[0.445 \pm 1.96 \cdot 0.016] = [0.414, 0.476]$. Because of the large sample and the fact that the distribution of θ is concentrated away from zero and one, the normal approximation works well in this example.

As already noted, when estimating a proportion, the normal approximation is generally improved by applying it to the logit transform, $\log(\frac{\theta}{1-\theta})$, which transforms the parameter space from the unit interval to the real line. The second histogram in Figure 2.3 shows the distribution of the transformed draws. The estimated posterior mean and standard deviation on the logit scale based on 1000 draws are -0.220 and 0.065. A normal approximation to the 95% posterior interval for θ is obtained by inverting the 95% interval on the logit scale $[-0.220 \pm 1.96 \cdot 0.065]$, which yields $[0.414, 0.477]$

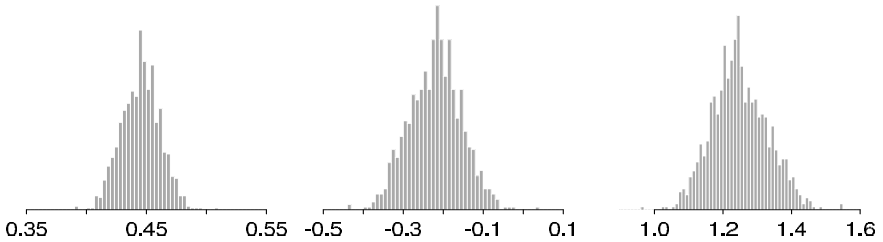

Figure 2.3 *Draws from the posterior distribution of (a) the probability of female birth, θ; (b) the logit transform, logit(θ); (c) the male-to-female sex ratio, $\phi = (1-\theta)/\theta$.*

Parameters of the prior distribution		Summaries of the posterior distribution	
$\frac{\alpha}{\alpha+\beta}$	$\alpha+\beta$	Posterior median of θ	95% posterior interval for θ
0.500	2	0.446	[0.415, 0.477]
0.485	2	0.446	[0.415, 0.477]
0.485	5	0.446	[0.415, 0.477]
0.485	10	0.446	[0.415, 0.477]
0.485	20	0.447	[0.416, 0.478]
0.485	100	0.450	[0.420, 0.479]
0.485	200	0.453	[0.424, 0.481]

Table 2.1 *Summaries of the posterior distribution of θ, the probability of a girl birth given placenta previa, under a variety of conjugate prior distributions.*

on the original scale. The improvement from using the logit scale is most noticeable when the sample size is small or the distribution of θ includes values near zero or one. In any real data analysis, it is important to keep the applied context in mind. The parameter of interest in this example is traditionally expressed as the 'sex ratio,' $(1-\theta)/\theta$, the ratio of male to female births. The posterior distribution of the ratio is illustrated in the third histogram. The posterior median of the sex ratio is 1.24, and the 95% posterior interval is [1.10, 1.41]. The posterior distribution is concentrated on values far above the usual European-race sex ratio of 1.06, implying that the probability of a female birth given placenta previa is less than in the general population.

Analysis using different conjugate prior distributions. The sensitivity of posterior inference about θ to the proposed prior distribution is exhibited in Table 2.1. The first row corresponds to the uniform prior distribution, $\alpha = 1$, $\beta = 1$, and subsequent rows of the table use prior distributions that are increasingly concentrated around 0.485, the proportion of female births in the general population. The first column shows the prior mean for θ, and the second column indexes the amount of prior information, as measured by $\alpha + \beta$; recall that $\alpha + \beta - 2$ is, in some sense, equivalent to the number of prior observations. Posterior inferences based on a large sample are not particularly sensitive to the prior distribution. Only at the bottom of the table, where the prior distribution contains information equivalent to 100 or 200 births, are the posterior intervals pulled noticeably toward the prior distribution, and even then, the 95% posterior intervals still exclude the prior mean.

Analysis using a nonconjugate prior distribution. As an alternative to the conjugate beta family for this problem, we might prefer a prior distribution that is centered around 0.485 but is flat far away from this value to admit the possibility that the truth is far away. The piecewise linear prior density in Figure 2.4a is an example

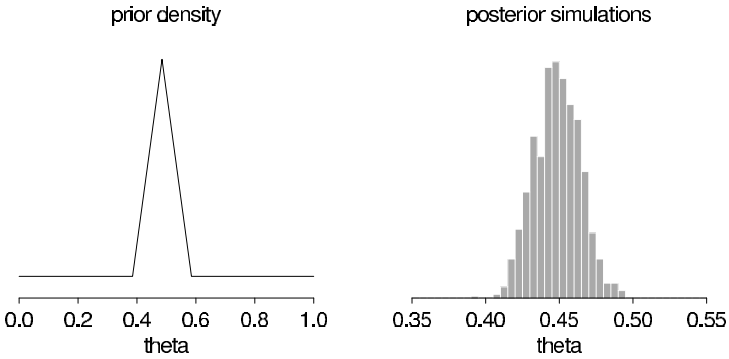

Figure 2.4 *(a) Prior density for θ in an example nonconjugate analysis of birth ratio example; (b) histogram of 1000 draws from a discrete approximation to the posterior density. Figures are plotted on different scales.*

of a prior distribution of this form; 40% of the probability mass is outside the interval $[0.385, 0.585]$. This prior distribution has mean 0.493 and standard deviation 0.21, similar to the standard deviation of a beta distribution with $\alpha+\beta = 5$. The unnormalized posterior distribution is obtained at a grid of θ values, $(0.000, 0.001, \ldots, 1.000)$, by multiplying the prior density and the binomial likelihood at each point. Posterior simulations can be obtained by normalizing the distribution on the discrete grid of θ values. Figure 2.4b is a histogram of 1000 draws from the discrete posterior distribution. The posterior median is 0.448, and the 95% central posterior interval is $[0.419, 0.480]$. Because the prior distribution is overwhelmed by the data, these results match those in Table 2.1 based on beta distributions. In taking the grid approach, it is important to avoid grids that are too coarse and distort a significant portion of the posterior mass.

2.5 Estimating a normal mean with known variance

The normal distribution is fundamental to most statistical modeling. The central limit theorem helps to justify using the normal likelihood in many statistical problems, as an approximation to a less analytically convenient actual likelihood. Also, as we shall see in later chapters, even when the normal distribution does not itself provide a good model fit, it can be useful as a component of a more complicated model involving t or finite mixture distributions. For now, we simply work through the Bayesian results assuming the normal model is appropriate. We derive results first for a single data point and then for the general case of many data points.

Likelihood of one data point

As the simplest first case, consider a single scalar observation y from a normal distribution parameterized by a mean θ and variance σ^2, where for this initial development we assume that σ^2 is known. The sampling distribution is

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2}.$$