.975 quantiles of the distribution. The distribution of $\theta^* - \hat{\theta}$ is approximated by subtracting $\hat{\theta} = .425$ from each θ_i^*, so the .025 and .975 quantiles of this distribution are estimated as

$$\underline{\delta} = .403 - .425 = -.022$$
$$\overline{\delta} = .446 - .425 = .021$$

Thus our approximate 95% confidence interval is

$$(\hat{\theta} - \overline{\delta}, \hat{\theta} - \underline{\delta}) = (.404, .447)$$

Since the uncertainty in $\hat{\theta}$ is in the second decimal place, this interval and that found in Example C are identical for all practical purposes. ■

EXAMPLE **E** Finally, we apply the bootstrap to find approximate confidence intervals for the parameters of the gamma distribution fit in Example C of Section 8.5. Recall that the estimates were $\hat{\alpha} = .471$ and $\hat{\lambda} = 1.97$. Of the 1000 bootstrap values of $\alpha^*, \alpha_1^*, \alpha_2^*, \ldots, \alpha_{1000}^*$, the 50th largest was .419 and the 950th largest was .538; the .05 and .95 quantiles of the distribution of $\alpha^* - \hat{\alpha}$ are approximated by subtracting $\hat{\alpha}$ from these values, giving

$$\underline{\delta} = .419 - .471 = -.052$$
$$\overline{\delta} = .538 - .471 = .067$$

Our approximate 90% confidence interval for α_0 is thus

$$(\hat{\alpha} - \overline{\delta}, \hat{\alpha} - \underline{\delta}) = (.404, .523)$$

The 50th and 950th largest values of λ^* were 1.619 and 2.478, and the corresponding approximate 90% confidence interval for λ_0 is (1.462, 2.321). ■

We caution the reader that there are a number of different methods of using the bootstrap to find approximate confidence intervals. We have chosen to present the preceding method largely because the reasoning leading to its development is fairly direct. Another popular method, the *bootstrap percentile method,* uses the quantiles of the bootstrap distribution of $\hat{\theta}$ directly. Using this method in the previous example, the confidence interval for α would be (.419, .538). Although this direct equation of quantiles of the bootstrap sampling distribution with confidence limits may seem initially appealing, its rationale is somewhat obscure. If the bootstrap distribution is symmetric, the two methods are equivalent (see Problem 38).

8.6 The Bayesian Approach to Parameter Estimation

A preview of the Bayesian approach was given in Example E of Section 3.5.2, which should be reviewed before continuing.

In the Bayesian approach, the unknown parameter θ is treated as a random variable, with "prior distribution" $f_\Theta(\theta)$ representing what we know about the parameter before observing data, X. In the following, we assume Θ is a continuous random variable; the discrete case is entirely analogous. This model is in contrast to the approaches described in the previous sections, in which θ was treated as an unknown constant. For a given value, $\Theta = \theta$, the data have the probability distribution (density or probability mass function) $f_{X|\Theta}(x|\theta)$. The joint distribution of X and Θ is thus

$$f_{X,\Theta}(x, \theta) = f_{X|\Theta}(x|\theta) f_\Theta(\theta)$$

and the marginal distribution of X is

$$\begin{aligned} f_X(x) &= \int f_{X,\Theta}(x, \theta) d\theta \\ &= \int f_{X|\Theta}(x|\theta) f_\Theta(\theta) d\theta \end{aligned}$$

The distribution of Θ given the data X is thus

$$\begin{aligned} f_{\Theta|X}(\theta|x) &= \frac{f_{X,\Theta}(x, \theta)}{f_X(x)} \\ &= \frac{f_{X|\Theta}(x|\theta) f_\Theta(\theta)}{\int f_{X|\Theta}(x|\theta) f_\Theta(\theta) d\theta} \end{aligned}$$

This is called the **posterior distribution;** it represents what is known about Θ having observed data X. Note that the likelihood is $f_{X|\Theta}(x|\theta)$, viewed as a function of θ, and we may usefully summarize the preceding result as

$$\begin{aligned} f_{\Theta|X}(\theta|x) &\propto f_{X|\Theta}(x|\theta) \times f_\Theta(\theta) \\ \text{Posterior density} &\propto \text{Likelihood} \times \text{Prior density} \end{aligned}$$

The Bayes paradigm has an appealing formal simplicity as it involves elementary probability operations. We will now see what it amounts to for examples we considered earlier.

EXAMPLE **A** *Fitting a Poisson Distribution*

Here the unknown parameter is λ, which has a prior distribution $f_\Lambda(\lambda)$, and the data are n i.i.d. observations $X_1, X_2, \dots, X_n$, which for a given value λ are Poisson random variables with

$$f_{X_i|\Lambda}(x_i|\lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}, \qquad x_i = 0, 1, 2, \dots$$

Their joint distribution given λ is (from independence) the product of their marginal distributions given λ

$$f_{X|\Lambda}(x|\lambda) = \frac{\lambda^{\Sigma_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!}$$

where X denotes $(X_1, X_2, \ldots, X_n)$. The posterior distribution of Λ given X is then

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda^{\Sigma_{i=1}^n x_i} e^{-n\lambda} f_\Lambda(\lambda)}{\int \lambda^{\Sigma_{i=1}^n x_i} e^{-n\lambda} f_\Lambda(\lambda)\, d\lambda}$$

(the term $\prod_{i=1}^n x_i!$ has cancelled out).

Thus, to evaluate the posterior distribution, we have to do two things: specify the prior distribution $f_\Lambda(\lambda)$ and carry out the integration in the denominator of the preceding expression. For illustration, we consider the data of Examples 8.4A and 8.5A.

We will consider two approaches to specifying the prior distribution. The first is that of an orthodox Bayesian who takes very seriously the model that the prior distribution specifies his prior opinion. Note that this specification should be done *before* seeing the data, X, and he is required to provide the probability density $f_\Lambda(\lambda)$ through introspection. This is not an easy task to carry out, and even the orthodox often compromise for convenience. He thus decides to quantify his opinion by specifying a prior mean $\mu_1 = 15$ and standard deviation $\sigma = 5$ and to use, because the math works out nicely as we will see, a gamma density with that mean and standard deviation. This choice could be aided by plotting gamma densities for various parameter values. The prior density is shown in Figure 8.9. Using the relationships developed in Example C in Section 8.4, the second moment is $\mu_2 = \mu_1^2 + \sigma^2 = 250$ and the parameters of the gamma density are

$$\nu = \frac{\mu_1}{\mu_2 - \mu_1^2} = 0.6$$

$$\alpha = \nu\mu_1 = 9$$

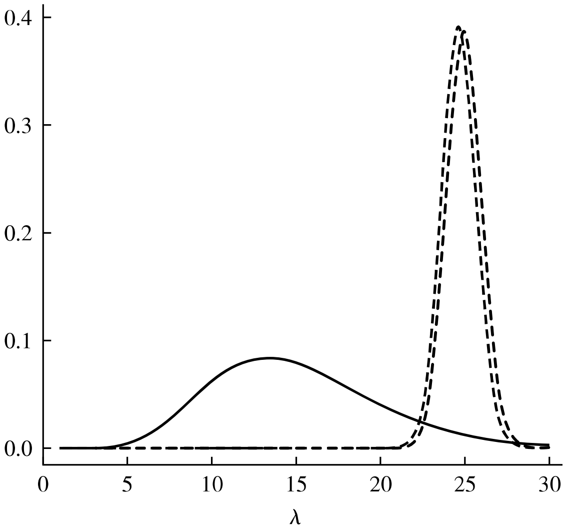

FIGURE 8.9 First statistician's prior (solid) and posterior (dashed). Second statistician's posterior (dotted).

(We denote the parameter by ν rather than by the usual λ since λ has already been used for the parameter of the Poisson distribution.) The prior distribution for Λ is then

$$f_\Lambda(\lambda) = \frac{\nu^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\nu\lambda}$$

After some cancellation, the posterior density is

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda^{\Sigma x_i+\alpha-1}e^{-(n+\nu)\lambda}}{\int_0^\infty \lambda^{\Sigma x_i+\alpha-1}e^{-(n+\nu)\lambda}d\lambda}$$

Now, consider this an important trick that is used time and again in Bayesian calculations: the denominator is a constant that makes the expression integrate to 1. We can deduce from the form of the numerator that the ratio *must* be a gamma density with parameters

$$\alpha' = \sum x_i + \alpha = 582$$
$$\nu' = n + \nu = 23.6$$

This standard trick allows the statistician to avoid having to do any explicit integration. (Make sure you understand it, because it will occur again several times.) The posterior density is shown in Figure 8.9. Compare it to the prior distribution to observe how observation of the data, X, has drastically changed his state of knowledge about Λ. Notice that the posterior density is much more symmetric and looks like a normal density (that this is no accident will be shown later). ■

According to the Bayesian paradigm, all the information about Λ is contained in the posterior distribution. The mean of this distribution (the **posterior mean**) is

$$\mu_{\text{post}} = \frac{\alpha'}{\nu'} = 24.7$$

The most probable value of Λ, the **posterior mode,** is 24.6. (Verify that the gamma density is maximized at $(\alpha - 1)/\nu$.) Either of these two values could be used as a point estimate of the unknown mean of the Poisson distribution, if a single number is required.

The variance of the posterior distribution is

$$\sigma^2_{\text{post}} = \frac{\alpha'}{\nu'^2} = 1.04$$

and the posterior standard deviation is $\sigma_{\text{post}} = 1.02$, which is a simple measure of variability—the posterior distribution of Λ has mean 24.7 and standard deviation 1.02. A Bayesian analogue of a 90% confidence interval is the interval from the 5th percentile to the 95th percentile of the posterior, which can be found numerically to be [23.02, 26.34]. A common alternative to this interval is a **high posterior density (HPD) interval,** formed as follows: Imagine placing a horizontal line at the high point of the posterior density and moving it downward until the interval of λ formed below where the line cuts the density contained 90% probability. If the posterior density is symmetric and unimodal, as is nearly the case in Figure 8.9, the HPD interval will coincide with the interval between the percentiles.

The second statistician takes a more utilitarian, noncommittal approach. She believes that it is implausible that the mean count λ could be larger than 100, and uses a simple prior that is uniform on [0, 100], without trying to quantify her opinion more precisely. The posterior density is thus

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda^{\Sigma_{i=1}^{n} x_i} e^{-n\lambda} \frac{1}{100}}{\frac{1}{100}\int_0^{100} \lambda^{\Sigma_{i=1}^{n} x_i} e^{-n\lambda} d\lambda}, \qquad 0 \leq \lambda \leq 100$$

The denominator has to be integrated numerically, but this is easy to do for such a smooth function. The resulting posterior density is shown in Figure 8.9. Using numerical evaluations, she finds that the posterior mode is 24.9, the posterior mean is 25.0, and the posterior standard deviation is 1.04. The interval from the 5th to the 95th percentile is [23.3, 26.7].

We now compare these two results to each other and to the results of maximum likelihood analysis.

Estimate	Bayes 1	Bayes 2	Maximum Likelihood
mode	24.6	24.9	24.9
mean	24.7	25.0	—
standard deviation	1.02	1.04	1.04
upper limit	26.3	26.7	26.6
lower limit	23.0	23.3	23.2

Comparing the results of the second Bayesian to those of maximum likelihood, it is important to realize that her posterior density is directly proportional to the likelihood for $0 \leq \lambda \leq 100$, because the prior is flat over this range and the posterior is proportional to the prior times the likelihood. Thus, her posterior mode and the maximum likelihood estimate are identical. There is no such guarantee that her posterior standard deviation and the approximate standard error of the maximum likelihood estimate are identical, but they turn out to be, to the number of significant figures displayed in the table. The two 90% intervals are very close.

Now compare the results of the first and second Bayesians. Observe that although his prior opinion was not in accord with the data, the data strongly modified the prior, to produce a posterior that is close to hers. Even though they start with quite different assumptions, the data forces them to very similar conclusions. His prior opinion has indeed influenced the results: his posterior mean and mode are less than hers, but the influence has been mild. (If there had been less data or if his prior opinions had been much more biased to low values, the results would have been in greater conflict.) The fundamental result that the posterior is proportional to the prior times the likelihood helps us to understand the difference: the likelihood is substantial only in the region approximately between $\lambda = 22$ and $\lambda = 28$. (This can be seen in the figure, because the second statistician's posterior is proportional to the likelihood. See Figure 8.5, also). In this region, his prior decreases slowly, so the posterior is proportional to a weighted version of the likelihood, with slowly decreasing weight.

The first Bayesian's posterior thus differs from the second by being pushed up slightly on the left and pulled down on the right.

Although they are very similar numerically, there is an important difference between the Bayesian and frequentist interpretation of the confidence intervals. In the Bayesian framework, Λ is a random variable and it makes perfect sense to say, "Given the observations, the probability that Λ is in the interval [23.3, 26.7] is 0.90." Under the frequentist framework, such a statement makes no sense, because λ is a constant, albeit unknown, and it either lies in the interval [23.3, 26.7] or doesn't—no probability is involved. Before the data are observed, the interval is random, and it makes sense to state that the probability that the interval contains the true parameter value is 0.90, but after the data are observed, nothing is random anymore. One way to understand the difference of interpretation is to realize that in the Bayesian analysis the interval refers to the state of knowledge about λ and not to λ itself.

Finally, we note that an alternative for the second statistician would have been to use a gamma prior because of its analytical convenience, but to make the prior very flat. This can be accomplished by setting α and λ to be very small.

EXAMPLE **B** *Normal Distribution*

It is convenient to reparametrize the normal distribution, replacing σ^2 by $\xi = 1/\sigma^2$; ξ is called the **precision.** We will also use θ in place of μ. The density is then

$$f(x|\theta, \xi) = \left(\frac{\xi}{2\pi}\right)^{1/2} \exp\left(-\frac{1}{2}\xi(x-\theta)^2\right)$$

The normal distribution has two parameters, and we will consider cases of Bayesian analysis depending on which of them are known and unknown. ■

Case of Unknown Mean and Known Variance

We first consider the case in which the precision is known, $\xi = \xi_0$ and the mean, θ, is unknown. In the Bayesian treatment, the mean is a random variable, Θ. It is mathematically convenient to use a prior distribution for Θ, which is $N(\theta_0, \xi_{\text{prior}}^{-1})$. This prior is very flat, or uninformative, when ξ_{prior} is very small, i.e., when the prior variance is very large. Thus, if $X = (X_1, X_2, \ldots, X_n)$ are independent given θ

$$\begin{aligned} f_{\Theta|X}(\theta|x) &\propto f_{X|\Theta}(x|\theta) \times f_\Theta(\theta) \\ &= \left(\frac{\xi_0}{2\pi}\right)^{n/2} \prod_{i=1}^{n} \exp\left(\frac{-\xi_0}{2}(x_i-\theta)^2\right) \times \left(\frac{\xi_{\text{prior}}}{2\pi}\right)^{1/2} \\ &\quad \times \exp\left(\frac{-\xi_{\text{prior}}}{2}(\theta-\theta_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\xi_0 \sum_{i=1}^{n}(x_i-\theta)^2 + \xi_{\text{prior}}(\theta-\theta_0)^2\right]\right) \end{aligned}$$

Here we have exhibited only the terms in the posterior density that depend upon θ; the last expression above shows the shape of the posterior density as a function of θ. The posterior density itself is proportional to this expression, with a proportionality constant that is determined by the requirement that the posterior density integrates to 1.

We will now manipulate the expression for the numerator to cast it in a form so that we can recognize that the posterior density is normal. Expressing $\sum(x_i - \theta)^2 = \sum(x_i - \bar{x})^2 + n(\theta - \bar{x})^2$, and absorbing more terms that do not depend on θ into the constant of proportionality (a typical move in Bayesian calculations), we find

$$f_{\Theta|X}(\theta|x) \propto \exp\left(-\frac{1}{2}[n\xi_0(\theta - \bar{x})^2 + \xi_{\text{prior}}(\theta - \theta_0)^2]\right)$$

Now, observe that this is of the form $\exp(-(1/2)Q(\theta))$, where $Q(\theta)$ is a quadratic polynomial. We can find expressions ξ_{post} and θ_{post}, and write

$$Q(\theta) = \xi_{\text{post}}(\theta - \theta_{\text{post}})^2 + \text{terms that do not depend on } \theta$$

and conclude that the posterior density is normal with posterior mean θ_{post} and posterior precision ξ_{post}. Again, terms that do not depend on θ do not affect the shape of the posterior density and are absorbed in the normalization constant that makes the posterior density integrate to 1. Thus we expand $Q(\theta)$ and identify the coefficient of θ^2 as the posterior precision and the coefficient of $-\theta$ as twice the posterior mean times the posterior precision. Doing so, we find

$$\xi_{\text{post}} = n\xi_0 + \xi_{\text{prior}}$$

$$\begin{aligned}\theta_{\text{post}} &= \frac{n\xi_0\bar{x} + \theta_0\xi_{\text{prior}}}{n\xi_0 + \xi_{\text{prior}}} \\ &= \bar{x}\frac{n\xi_0}{n\xi_0 + \xi_{\text{prior}}} + \theta_0\frac{\xi_{\text{prior}}}{n\xi_0 + \xi_{\text{prior}}}\end{aligned}$$

The posterior density of θ is thus normal with this mean and precision. Note that the precision has increased and that the posterior mean is a weighted combination of the sample mean and the prior mean.

To interpret these results, consider what happens when $\xi_{\text{prior}} \ll n\xi_0$, which would be the case if n were sufficiently large of if ξ_{prior} were small (as for a very flat prior). Then the posterior mean would be

$$\theta_{\text{post}} \approx \bar{x}$$

which is the maximum likelihood estimate, and

$$\xi_{\text{post}} \approx n\xi_0$$

This last equation can be written as $\sigma_{\text{post}}^2 = \sigma_0^2/n$, which is just the variance of $\overline{X}$ in the non-Bayesian setting. In summary, if the flat prior with very small ξ_{prior} is used, the posterior density of θ is very close to normal with mean $\bar{x}$ and variance σ_0^2/n. ■

Case of Known Mean and Unknown Variance

In this case, the precision is unknown and is treated as a random variable Ξ, with prior distribution $f_\Xi(\xi)$. Given ξ, the X_i are independent $N(\theta_0, \xi^{-1})$. Let $X = (X_1, X_2, \ldots, X_n)$. Then

$$\begin{aligned}f_{\Xi|X}(\xi|x) &\propto f_{X|\Xi}(x|\xi)f_\Xi(\xi) \\ &\propto \xi^{n/2}\exp\left(-\frac{1}{2}\xi\sum(x_i - \theta_0)^2\right) f_\Xi(\xi)\end{aligned}$$

Observing how the density depends on ξ, we realize that it is analytically convenient to specify the prior to be a gamma density: $\Xi \sim \Gamma(\alpha, \lambda)$. Then

$$f_{\Xi|X}(\xi|x) \propto \xi^{n/2} \exp\left(-\frac{1}{2}\xi \sum (x_i - \theta_0)^2\right) \xi^{\alpha-1} e^{-\lambda\xi}$$

which is a gamma density with parameters,

$$\alpha_{\text{post}} = \alpha + \frac{n}{2}$$
$$\lambda_{\text{post}} = \lambda + \frac{1}{2}\sum (x_i - \theta_0)^2$$

In the case of a flat prior (small α and λ), the posterior mean and mode are

$$\text{Posterior mean} \approx \frac{1}{n}\sum (x_i - \theta_0)^2$$
$$\text{Posterior mode} \approx \frac{1}{n-2}\sum (x_i - \theta_0)^2$$

The former is the maximum likelihood estimate of σ^2. In the limit, $\lambda \to 0$, $\alpha \to 0$,

$$f_{\Xi|X}(\xi|x) \propto \xi^{n/2-1} \exp\left(-\frac{1}{2}\xi \sum (x_i - \theta_0)^2\right) \qquad \blacksquare$$

Case of Unknown Mean and Unknown Variance

In this case, there are two unknown parameters, and a Bayesian approach requires the specification of a joint two-dimensional prior distribution. We follow a path of mathematical convenience and take the priors to be independent:

$$\Theta \sim N\left(\theta_0, \xi_{\text{prior}}^{-1}\right)$$
$$\Xi \sim \Gamma(\alpha, \lambda)$$

We then have

$$\begin{aligned} f_{\Theta,\Xi|X}(\theta, \xi|x) &\propto f_{X|\Theta,\Xi}(x|\theta, \xi) f_\Theta(\theta) f_\Xi(\xi) \\ &\propto \xi^{n/2} \exp\left(-\frac{\xi}{2}\sum (x_i - \theta)^2\right) \\ &\quad \times \exp\left(-\frac{\xi_{\text{prior}}}{2}(\theta - \theta_0)^2\right) \xi^{\alpha-1} \exp(-\lambda\xi) \end{aligned}$$

From the manner in which θ and ξ occur in the first exponential, it appears that the two variables are not independent in the posterior even though they were in the prior. To evaluate this joint posterior density, we would have to find the constant of proportionality that makes it integrate to 1—the normalization constant. Two dimensional numerical integration could be used.

Often the primary interest is in the mean, θ, and one useful aspect of Bayesian analysis is that information about θ can be "marginalized" by integrating out ξ:

$$f_{\Theta|X}(\theta|x) = \int_0^\infty f_{\Theta,\Xi|X}(\theta, \xi|x) d\xi$$

Examining the preceding expression for $f_{\Theta,\Xi|X}(\theta,\xi|x)$ as a function of ξ, we see that it is of the form of a gamma density, with parameters $\tilde{\alpha} = \alpha + n/2$ and $\tilde{\lambda} = \lambda + (1/2)\sum(x_i - \theta)^2$, so we can evaluate the integral. We thus find

$$f_{\Theta|X}(\theta|x) \propto \exp\left(-\frac{\xi_{\text{prior}}}{2}(\theta-\theta_0)^2)\right) \frac{\Gamma(\alpha+n/2)}{[\lambda + \frac{1}{2}\sum(x_i-\theta)^2]^{\alpha+n/2}}$$

This is not a density that we recognize, but it could be evaluated numerically. Doing so would again entail finding the normalizing constant, which could be done by numerical integration. Some simplifications occur when n is large or when the prior is quite flat (α, λ, ξ_{prior} are small). Then

$$f_{\Theta|X}(\theta|x) \propto \left(\sum(x_i-\theta)^2\right)^{-n/2}$$

This posterior is maximized when $\sum(x_i-\theta)^2$ is minimized, which occurs at $\theta = \bar{x}$. We can relate this to the result we found for maximum likelihood analysis by expressing

$$\begin{aligned}\sum(x_i-\theta)^2 &= \sum(x_i-\bar{x})^2 + n(\theta-\bar{x})^2\\ &= (n-1)s^2 + n(\theta-\bar{x})^2\\ &= (n-1)s^2\left(1+\frac{n(\theta-\bar{x})^2}{(n-1)s^2}\right)\end{aligned}$$

Substituting this above and absorbing terms that do not depend on θ into the proportionality constant, we find

$$f_{\Theta|X}(\theta|x) \propto \left(1+\frac{1}{n-1}\frac{n(\theta-\bar{x})^2}{s^2}\right)^{-n/2}$$

Now comparing this to the definition of the t distribution (Section 6.2), we see that

$$\frac{\sqrt{n}(\Theta-\bar{x})}{s} \sim t_{n-1}$$

corresponding to the result from maximum likelihood analysis.

The interval $\bar{x} \pm t_{n-1}(\alpha/2)s/\sqrt{n}$ was earlier derived as a $100(1-\alpha)\%$ confidence interval centered about the maximum likelihood estimate, and here it has reappeared in the Bayesian analysis as an interval with posterior probability $1-\alpha$. There are differences of interpretation, however, just as there were for the earlier Poisson case. The Bayesian interval is a probability statement referring to the state of knowledge about θ given the observed data, regarding θ as a random variable. The frequentist confidence interval is based on a probability statement about the possible values of the observations, regarding θ as a constant, albeit unknown. ■

EXAMPLE **C** *Hardy-Weinberg Equilibrium*

We now turn to a Bayesian treatment of Example A in Section 8.5.1. We use the multinomial likelihood function and a prior for θ, which is uniform on [0, 1]. The posterior density is thus proportional to the likelihood, and is shown in Figure 8.10. Note that it looks very much like a normal density, a phenomenon that will be explored in a later section. Since $f_{X|\Theta}(x|\theta)$ is a polynomial in θ (of high degree),

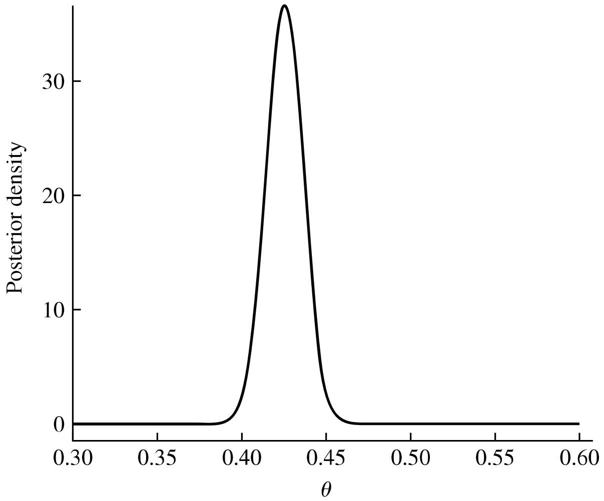

FIGURE **8.10** Posterior distribution of Θ.

the normalization constant can in principle be computed analytically. (Alternatively, all the computations can be done numerically.)

Because the prior is flat, the posterior is directly proportional to the likelihood and the maximum of the posterior density is the maximum likelihood estimate, $\hat{\theta} = 0.4247$. The 0.025 percentile of the density is 0.404, and the 0.975 percentile is 0.446. These results agree with the approximate confidence interval found for the maximum likelihood estimate in Example C in Section 8.5.3. ■

8.6.1 Further Remarks on Priors

In the previous section, we saw that if the prior for a Poisson parameter is chosen to be a gamma density, then the posterior is also a gamma density. Similarly, when the prior for a normal mean with known variance is chosen to be normal, then the posterior is normal as well. Earlier, in Example E in Section 3.5.2, a beta prior was used for a binomial parameter, and the posterior turned out to be beta as well. These are examples of **conjugate priors:** if the prior distribution belongs to a family G and, conditional on the parameters of G, the data have a distribution H, then G is said to be conjugate to H if the posterior is in the family G. Other conjugate priors will be the subject of problems at the end of the chapter. Conjugate priors are used for mathematical convenience (required integrations can be done in closed form) and because they can assume a variety of shapes as the parameters of the prior are varied.

In scientific applications, it is usually desirable to use a flat, or "uninformative," prior so that the data can speak for themselves. Even if a scientific investigator actually had a strong prior opinion, he or she might want to present an "objective" analysis. This is accomplished by using a flat prior so that the conclusions, as summarized in the posterior density, are those of one who is initially unopinionated or unprejudiced.

If an informative prior were used, it would have to be justified to the larger scientific community. The objective prior thus has a hypothetical, or "what if," status: if one was initially indifferent to parameter values in the range in which the likelihood is large, then one's opinion after observing the data would be expressed as a posterior proportional to the likelihood.

Attempts have been made to formalize more precisely what the notion of an uninformative prior means. One problem that is addressed is caused by reparametrization. For example, suppose that the prior density of the precision ξ is taken to be uniform on an interval $[a, b]$, which might seem to be a reasonable way to quantify the notion of being uniformative. However, if the variance $\sigma^2 = 1/\xi$, rather than the precison, was used, the prior density of σ^2 would not be uniform on $[b^{-1}, a^{-1}]$. We will not delve further into these issues here, except to note that the parametrization θ or $g(\theta)$ would make a difference only if the difference in the shapes of the priors was substantial in the region in which the likelihood was large.

We saw in the Poisson example that if α and ν are very small, the gamma prior is quite flat and the posterior is proportional to the likelihood function. Formally, if α and ν are set equal to zero, then the prior is

$$f_{\Lambda|\alpha,\nu}(\lambda) = \lambda^{-1}, \quad 0 \leq \lambda < \infty$$

But this function does not integrate to 1—it is not a probability density. A similar phenomena occurs in the normal case with unknown mean and known precision, if the prior precision is set equal to 0. The prior is then

$$f_\Theta(\theta) \propto 1, \quad -\infty < \theta < \infty$$

and not a probability density either. Such priors are called **improper priors** (priors that lack propriety).

In general, if an improper prior is formally used, the posterior may not be a density either, because the denominator of the expression for the posterior density, $\int f_{X|\Theta}(x|\theta) f_\Theta(\theta)\, d\theta$ may not converge. (Note that it is integrated with respect to θ, not x.) This has not been the case in our examples. For the Poisson example, if $f_\Lambda(\lambda) \propto \lambda^{-1}$, then the denominator is

$$\int_0^\infty \lambda^{\sum x_i - 1} e^{-n\lambda} d\lambda < \infty$$

In the normal case, too, the integral is defined, and thus there is a well-defined posterior density.

Let us revisit some examples using the device of an improper prior. In the Poisson example, using the improper prior $f_\Lambda(\lambda) = \lambda^{-1}$ results in a (proper) posterior

$$f_{\Lambda|X}(\lambda|x) \propto \lambda^{\sum x_i - 1} e^{-n\lambda}$$

which can be recognized as a gamma density.

In the normal example with unknown mean and variance, we can take θ and ξ to be independent with improper priors $f_\Theta(\theta) = 1$ and $f_\Xi(\xi) = \xi^{-1}$. The joint posterior of θ and ξ is then

$$f_{\Theta,\Xi|X}(\theta, \xi|x) \propto \xi^{n/2-1} \exp\left(-\frac{\xi}{2} \sum (x_i - \theta)^2\right)$$

Expressing $\sum_{i=1}^{n}(x_i - \theta)^2 = (n-1)s^2 + n(\theta - \bar{x})^2$, we have

$$f_{\Theta,\Xi|X}(\theta, \xi|x) \propto \xi^{n/2-1} \exp\left(-\frac{\xi}{2}(n-1)s^2\right) \exp\left(-\frac{n\xi}{2}(\theta - \bar{x})^2\right)$$

For fixed ξ, this expression is proportional to the conditional density of θ given ξ. (Why?) From the form of the dependence on θ, we see that conditional on ξ, θ is normal with mean $\bar{x}$ and precision $n\xi$. By integrating out ξ, we can find the marginal distribution of θ and relate it to the t distribution as was done earlier.

Since improper priors are not actually probability densities, they are difficult to interpret literally. However, the resulting posteriors can be viewed as approximations to those that would have arisen with extreme values of the parameters of proper priors. The priors corresponding to such extreme values are very flat, so the posterior is dominated by the likelihood. Then it is only in the range in which the likelihood is large that the prior makes any practical difference—truncating the improper prior well outside this range to produce a proper prior will not appreciably change the posterior.

8.6.2 Large Sample Normal Approximation to the Posterior

We have seen in several examples that the posterior distribution is nearly normal with the mean equal to the maximum likelihood estimate, and that the posterior standard deviation is close to the asymptotic standard deviation of the maximum likelihood estimate. The two methods thus often give quite comparable results. We will not give a formal proof here, but rather will sketch an argument that the posterior distribution is approximately normal with the mean equal the the maximum likelihood estimate, $\hat{\theta}$, and variance approximately equal to $-[l''(\hat{\theta})]^{-1}$.

Denoting the observations generically by x, the posterior distribution is

$$\begin{aligned} f_{\Theta|X}(\theta|x) &\propto f_\Theta(\theta) f_{X|\Theta}(x|\theta) \\ &= \exp[\log f_\Theta(\theta)] \exp[\log f_{X|\Theta}(x|\theta)] \\ &= \exp[\log f_\Theta(\theta)] \exp[l(\theta)] \end{aligned}$$

Now, if the sample is large, the posterior is dominated by the likelihood, and in the region where the likelihood is large, the prior is nearly constant. Thus, to an approximation,

$$\begin{aligned} f_{\Theta|X}(\theta|x) &\propto \exp\left[l(\hat{\theta}) + (\theta - \hat{\theta})l'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 l''(\hat{\theta})\right] \\ &\propto \exp\left[\frac{1}{2}(\theta - \hat{\theta})^2 l''(\hat{\theta})\right] \end{aligned}$$

In the last step, we used the fact that since $\hat{\theta}$ is the maximum likelihood estimate $l'(\hat{\theta}) = 0$. The term $l(\hat{\theta})$ was absorbed into a proportionality constant, since we are evaluating the posterior as a function of θ. Finally, observe that the last expression is proportional to a normal density with mean $\hat{\theta}$ and variance $-[l''(\hat{\theta})]^{-1}$.