In order for the rejection method to be computationally efficient, the algorithm should lead to acceptance with high probability; otherwise, many rejection steps may have to be looped through for each acceptance.

EXAMPLE **E** *Bayesian Inference*

A freshly minted coin has a certain probability of coming up heads if it is spun on its edge, but that probability is not necessarily equal to $\frac{1}{2}$. Now suppose it is spun n times and comes up heads X times. What has been learned about the chance the coin comes up heads? We will go through a Bayesian treatment of this problem. Let Θ denote the probability that the coin will come up heads. We represent our knowledge about Θ before gathering any data by a probability density on [0, 1], called the **prior density.** If we are totally ignorant about Θ, we might represent our state of knowledge by a uniform density on [0, 1]:

$$f_\Theta(\theta) = 1, \quad 0 \le \theta \le 1.$$

We will see how observing X changes our knowledge about Θ, transforming the prior distribution into a "posterior" distribution.

Given a value θ, X follows a binomial distribution with n trials and probability of success θ:

$$f_{X|\Theta}(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad x = 0, 1, \ldots, n$$

Now Θ is continuous and X is discrete, and they have a joint probability distribution:

$$\begin{aligned} f_{\Theta,X}(\theta, x) &= f_{X|\Theta}(x|\theta)f_\Theta(\theta) \\ &= \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad x = 0, 1, \ldots, n, \quad 0 \le \theta \le 1 \end{aligned}$$

This is a density function in θ and a probability mass function in x, an object of a kind we have not seen before. We can calculate the marginal density X by integrating the joint over θ:

$$f_X(x) = \int_0^1 \binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta$$

We can calculate this formidable looking integral by a trick. First write

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} = \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)}$$

(If k is an integer, $\Gamma(k) = (k-1)!$; see Problem 49 in Chapter 2). Recall the beta density (Section 2.2.4)

$$g(u) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}u^{a-1}(1-u)^{b-1}, \qquad 0 \le u \le 1$$

The fact that this density integrates to 1 tells us that

$$\int_0^1 u^{a-1}(1-u)^{b-1}du = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Thus, identifying u with θ, $a-1$ with x, and $b-1$ with $n-x$,

$$\begin{aligned} f_X(x) &= \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} \int_0^1 \theta^x(1-\theta)^{n-x} d\theta \\ &= \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} \\ &= \frac{1}{n+1}, \qquad x = 0, 1, \ldots, n \end{aligned}$$

Thus, if our prior on θ is uniform, each outcome of X is *a priori* equally likely.

Our knowledge about Θ having observed $X = x$ is quantified in the conditional density of Θ given $X = x$:

$$\begin{aligned} f_{\Theta|X}(\theta|x) &= \frac{f_{\Theta,X}(\theta,x)}{f_X(x)} \\ &= (n+1)\binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= (n+1)\frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^x(1-\theta)^{n-x} \\ &= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^x(1-\theta)^{n-x} \end{aligned}$$

The relationship $x\Gamma(x) = \Gamma(x+1)$ has been used in the second step (see Problem 49, Chapter 2). Bear in mind that for each fixed x, this is a function of θ—the posterior density of θ given x—which quantifies our opinion about Θ having observed x heads in n spins. The posterior density is a beta density with parameters $a = x+1$, $b = n-x+1$.

A one-Euro coin has the number 1 on one face and a bird on the other face. I spun such a coin 20 times: the 1 came up 13 of the 20 times. Using the prior, $\Theta \sim U[0, 1]$, the posterior is beta with $a = x+1 = 14$ and $b = n-x+1 = 8$. Figure 3.16 shows this posterior, which represents my opinion if I was initially totally ignorant of θ and then observed thirteen 1s in 20 spins. From the figure, it is extremely unlikely that $\theta < 0.25$, for example. My probability, or belief, that θ is greater than $\frac{1}{2}$ is the area under the density to the right of $\frac{1}{2}$, which can be calculated to be 0.91. I can be 91% certain that θ is greater than $\frac{1}{2}$.

We need to distinguish between the steps of the preceding probability calculations, which are are mathematically straightforward; and the interpretation of the results, which goes beyond the mathematics and requires a model that belief can be expressed in terms of probability and revised using the laws of probability. See Figure 3.16. ■

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 $X_1, X_2, \ldots, X_n$, which for a given value λ are Poisson random variables with

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

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

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

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

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

(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

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

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

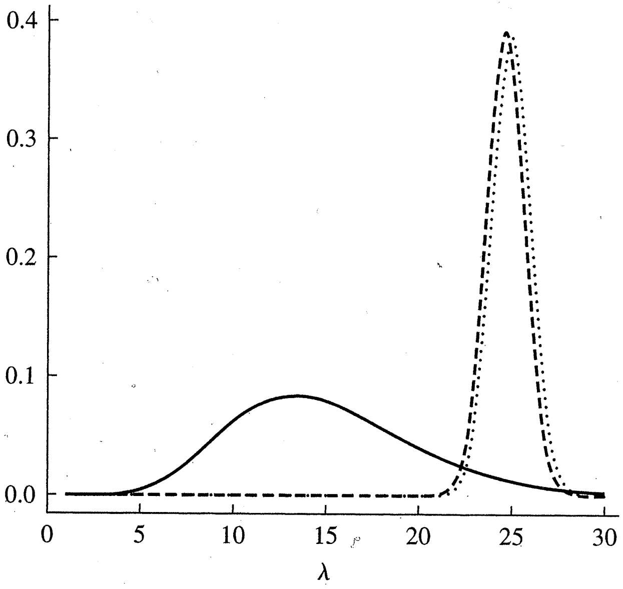

FIGURE 8.9 First statistician's prior (solid) and posterior (dashed). Second statistician's posterior (dotted).

(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

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

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 $0 \le \lambda \le 100$, 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.