Chapter 6

Bayesian Inference

6.1. BACKGROUND

The Bayesian, or subjective, view of probability leads to a framework for statistical inference that is different from the more familiar "frequentist" methods that are the subject of Chapter 5. Bayesian inference is parametric, in that the subjects of the inferences are the parameters of probability distributions, of the kinds described in Chapter 4. A parametric distribution is assumed in order to characterize quantitatively the nature of the data-generating process, and its mathematical dependence on the parameter(s) about which inferences are being drawn. For example, if the data at hand have resulted from N independent and identical Bernoulli trials, then it would be natural to adopt the binomial distribution (Equation 4.1) as the data-generating model. The target of statistical inference would then be the binomial parameter, p, and inferences about p could then be used to more fully characterize the nature of the data-generating process.

Regarding probability as a quantitative expression of subjective degree of belief leads to two distinctive differences between the structures of Bayesian and frequentist inference. The first is that prior information (i.e., information available before the current data have been obtained or seen) about the parameter(s) of interest, often reflecting the analyst's subjective judgment, is quantified by a probability distribution. This distribution may or may not be of a familiar parametric form, such as one of the distributions discussed in Chapter 4. The calculations underlying Bayesian inference combine this prior information with the information provided by the data, in an optimal way.

The second difference between the two modes of inference has to do with the ways in which the parameters that are the targets of inference are viewed. In the frequentist view, parameters of a data-generating model are fixed, if unknown, constants. Accordingly in this view it makes no sense to think about or try to characterize uncertainty about them, since they are unvarying. Rather, frequentist inferences about parameters are made on the basis of the distribution of data statistics under (possibly hypothetical) repeated sampling. In contrast, the Bayesian approach allows the parameter(s) being studied to be regarded as being subject to uncertainty that can be quantified using a probability distribution, which is derived by combining the prior information with the data, in light of the chosen data-generating model.

The relative merits of frequentist and Bayesian inference continue to be debated within the statistics profession. A summary of the recent state of these discussions is provided by Little (2006).

6.2. THE STRUCTURE OF BAYESIAN INFERENCE

6.2.1. Bayes' Theorem for Continuous Variables

The computational algorithm for Bayesian inference is provided by Bayes' Theorem, which was presented for discrete variables in Equation 2.16. However, even if the data upon which inferences will be based are discrete, the parameters that are the subject of inference are generally continuous, in which

Statistical Methods in the Atmospheric Sciences. https://doi.org/10.1016/B978-0-12-815823-4.00006-7

case the probability distributions characterizing their uncertainty (the analyst's degrees of belief) may be represented as probability density functions. Analogously to Equation 2.16, Bayes' Theorem for continuous probability models can be expressed as

$$f(\theta|\, x) = \frac{f(x|\, \theta) f(\theta)}{f(x)} = \frac{f(x|\, \theta) f(\theta)}{\int_{\theta} f(x|\, \theta) f(\theta)\, d\theta}. \tag{6.1}$$

Here θ represents the parameter(s) about which inferences are to be drawn (e.g., a binomial probability p or a Poisson rate μ), and x represents the available data.

Equation 6.1 expresses the optimal combination of prior information and the available data for inference regarding the parameter(s) θ. Prior subjective beliefs and/or objective information regarding θ is quantified by the *prior distribution*, $f(\theta)$, which will be a continuous PDF when θ is a continuous parameter. It may be nontrivial to make a good assessment of the prior distribution for a given problem, and different analysts may reasonably reach different conclusions regarding it. The impact of the prior distribution, and consequences of different choices for it for inferences regarding θ, will be presented in more detail in Section 6.2.3.

The general nature of the data-generating process, and the quantitative influence of different values of θ on it, are represented by the *likelihood*, $f(x\,|\,\theta)$. Notationally, the likelihood appears to be identical to the probability distribution function representing the data-generating process for discrete data or to the PDF for continuous data. However, the distinction is that the likelihood is a function of the parameter(s) θ for fixed values of the data x, as was the case in Section 4.6.1, rather than a function of the data for fixed parameters. The function $f(x\,|\,\theta)$ expresses the relative plausibility ("likelihood") of the data at hand as a function of (given) different possible values for θ.

If the data are discrete, then the likelihood will look notationally like the probability distribution function chosen to represent the data-generating process, for example, Equation 4.1 for the binomial distribution or Equation 4.12 for the Poisson distribution. Both of these likelihoods would be functions of a continuous variable, that is, $\theta = p$ for the binomial, and $\theta = \mu$ for the Poisson. However, the likelihood $f(x\,|\,\theta)$ is generally not a PDF. Even though (for discrete x) $\Sigma_x \Pr\{X = x\} = 1$, in general $\int_\theta f(x\,|\,\theta)\, d\theta \neq 1$. If the data x are continuous, so that the likelihood looks notationally like the data PDF, the likelihood will in general also be a continuous function of the parameter(s) θ, but will typically also not itself be a PDF, again because $\int_\theta f(x\,|\,\theta)\, d\theta \neq 1$ even though $\int_x f(x\,|\,\theta)\, dx = 1$.

The optimal combination of the prior information, $f(\theta)$, and the information provided by the data in the context of the assumed character of the data-generating process, $f(x\,|\,\theta)$, is achieved through the product in the numerator on the right-hand side of Equation 6.1. The result is the *posterior distribution*, $f(\theta\,|\,x)$, which is the PDF for the parameter(s) θ characterizing the current best information regarding uncertainty about θ. The posterior distribution results from the process of updating the prior distribution in light of the information provided by the data, as seen through model provided by the likelihood for representing the data-generating process.

For settings in which all the data do not become available at the same time, this Bayesian updating can be computed sequentially. In such cases the analyst's assessment of the parameter uncertainty in the prior distribution $f(\theta)$ is first updated using whatever data is available initially, to yield a first iteration of the posterior distribution. That posterior distribution can then be further updated as new data become available, by applying Bayes' theorem with that initially calculated posterior distribution now playing the role of the prior distribution. The result of iterating Bayes' theorem in this way will be identical to what would be obtained if all the data had been used at the same time for a single updating of the initial prior distribution.

In order for the posterior distribution to be a proper PDF (i.e., integrating to 1), the product of the likelihood and the prior is scaled by the value $f(x) = \int_\theta f(x|\theta) f(\theta)\, d\theta$ for the available data x, in the denominator of Equation 6.1. Because the important work of Equation 6.1 occurs in the numerator of the right-hand side, that equation is sometimes expressed simply as

$$f(\theta|x) \propto f(x|\theta) f(\theta), \tag{6.2}$$

or "the posterior is proportional to the likelihood times the prior."

Example 6.1. Iterative Use of Bayes' Theorem

Consider the simple but instructive situation in which data for the number of "successes" x in a sequence of N independent and identical Bernoulli trials are to be used to estimate the success probability for future trials. This parameter p controls the nature of the data-generating process in this setting, and clearly the relationship of the success probability to possible realizations of the data (i.e., the data-generating process) is provided by the binomial distribution (Equation 4.1). Accordingly the natural choice for the likelihood is

$$f(x|p) = \binom{N}{x} p^x (1-p)^{N-x} \propto p^x (1-p)^{N-x}, \tag{6.3}$$

where the success probability p is the parameter θ about which inferences are to be made. The proportionality indicated in the second part of Equation 6.3 is appropriate because the combinatorial part of the binomial probability distribution function does not involve p, and so will factor out of the integral in the denominator of Equation 6.1 and cancel, for any choice of the prior distribution $f(p)$. Equation 6.3 is notationally identical to the discrete probability distribution function for the binomial distribution, Equation 4.1. However, unlike Equation 4.1, Equation 6.3 is not a discrete function of x, but rather is a continuous function of p, for a fixed number of successes x over the course of N independent trials.

An appropriate prior distribution $f(p)$ characterizing an analyst's initial uncertainty regarding possible values for p will depend on what, if any, information about p might be available before new data will be observed, as will be discussed more fully in Section 6.2.3. However, since $0 \le p \le 1$, any reasonable choice for $f(p)$ will have support on this interval. If the analyst has no initial idea regarding which values of p might be more or less likely, a reasonable prior might be the uniform distribution, $f(p) = 1$ (Section 4.4.6), which expresses the judgment that no value of p on the interval $0 \le p \le 1$ seems initially more plausible than any other.

Suppose now that the results of $N = 10$ Bernoulli trials from a process of interest become available, and of these $x = 2$ are successes. Bayes' Theorem provides the recipe for updating the initial indifference among possible values for p that is expressed by $f(p) = 1$, in the light of the results of these $N = 10$ observations. According to Equation 6.1, the posterior distribution is

$$f(p|x) = \frac{\binom{10}{2} p^2 (1-p)^8 \bullet 1}{\binom{10}{2} \int_0^1 p^2 (1-p)^8 dp} = \frac{\Gamma(12)}{\Gamma(3)\,\Gamma(9)} p^2 (1-p)^8. \tag{6.4}$$

Alternatively, using Equation 6.2,

$$f(p|x) \propto p^2 (1-p)^8 \bullet 1, \tag{6.5}$$

which achieves the same result, because the integral in the denominator of Equation 6.4 yields $\Gamma(3)\Gamma(9)/\Gamma(12)$, which is exactly the factor required for the posterior distribution $f(p|x)$ to integrate to 1 over $0 \le p \le 1$ and thus to be a PDF. In this case the posterior distribution is a beta distribution (Equation 4.58), with parameters $\alpha = x+1 = 3$ and $\beta = N-x+1 = 9$. Posterior distributions will not always turn out to be recognizable and familiar parametric forms, but a beta distribution has resulted here because of the nature of the chosen prior distribution and its interaction with the specific mathematical form of the likelihood in Equation 6.3, as will be explained in Section 6.3.

The posterior distribution in Equation 6.4 is the result of updating the initial prior distribution in light of having observed $x = 2$ successes in $N = 10$ Bernoulli trials. It thus quantitatively expresses the degree of belief regarding the possible values for p after having observed these data, for an analyst whose prior beliefs had been well represented by $f(p) = 1$.

Consider now how these beliefs should change if data from additional realizations of the same Bernoulli process become available. Bayes' Theorem will be iterated again, updating the current state of knowledge or belief in light of the new data. The prior distribution for this next iteration of Bayes' Theorem is not the initial prior $f(p) = 1$, but rather the posterior distribution from the most recent probability updating, that is, the beta distribution from Equation 6.4. Suppose the next data observed are the results of $N = 5$ Bernoulli trials, of which $x = 3$ are successes. The second application of Equation 6.2 yields

$$f(p|x) \propto p^x(1-p)^{N-x}\, p^2(1-p)^8 = p^{x+2}(1-p)^{N-x+8} = p^5(1-p)^{10}. \tag{6.6}$$

Neither the combinatorial part of the likelihood in Equation 6.3 or the ratio of gamma functions in the new prior distribution (Equation 6.4) depend on p, and so both cancel in the quotient of Equation 6.1. This updated posterior distribution is also a beta distribution, now with parameters $\alpha = 6$ and $\beta = 11$.

Figure 6.1 illustrates this probability updating process, by comparing the initial prior distribution $f(p) = 1$; the first posterior distribution (the beta distribution, Equation 6.4), with $\alpha = 3$, and $\beta = 9$, which becomes the next prior distribution; and the final posterior distribution (the beta distribution with $\alpha = 6$,

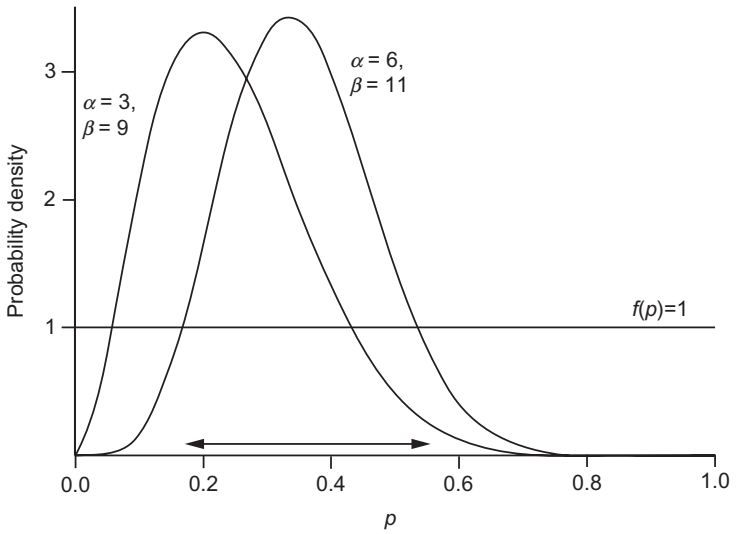

FIGURE 6.1 The prior distribution $f(p) = 1$, and two posterior beta distributions obtained after one ($\alpha = 3$, $\beta = 9$), and two ($\alpha = 6$, $\beta = 11$) applications of Equation 6.1, reflecting information contained in the two data installments. The double-headed arrow indicates the 95% CCI for p according to the second ($\alpha = 6$, $\beta = 11$) posterior distribution.

and $\beta = 11$, Equation 6.6). After the first application of Bayes' Theorem, it is evident that the most likely values for p are near the success relative frequency $x/N = 2/10$, and that values of $p > 0.7$ are associated with very small probability. After the second installment of data has been processed the most likely values for p are near the success relative frequency for all 15 realizations, or 5/15. If a single application of Bayes' Theorem had been made, using all of these data at once (i.e., $N = 15$ and $x = 5$), exactly the same posterior distribution (Equation 6.6) would have resulted from updating the original uniform prior distribution $f(p) = 1$. Similarly, if Equation 6.1 had been iterated 15 times, each using one of the Bernoulli realizations, the same posterior distribution would have resulted, regardless of the order in which the $x = 5$ successes and $N - x = 10$ nonsuccesses had been presented. ◇

6.2.2. Inference and the Posterior Distribution

The posterior distribution, $f(\theta \mid x)$, provides the basis for statistical inference in the Bayesian framework. It is the result, through application of Bayes' Theorem, of the combination of prior beliefs about θ with information about θ contained in the data x. Thus communication of the posterior density fully expresses the analyst's beliefs regarding θ. When the posterior distribution is of a conventional parametric form (e.g., the beta distributions in Example 6.1), quoting the parameters of the posterior distribution is a compact and convenient way to communicate the analyst's degree of belief and uncertainty regarding θ. The parameters of the posterior distribution (and also of the prior distribution) are known as *hyperparameters*, in order to more easily distinguish them from the parameter(s) that are the subjects of the statistical inference. In Example 6.1, inferences about the binomial parameter p were computed and expressed in terms of a beta posterior distribution, whose hyperparameters were $\alpha = 6$ and $\beta = 11$.

Especially if the posterior distribution is not of a familiar parametric form, for some purposes one might want to provide a point estimate for the parameter θ that is the subject of inference. There are several plausible choices for this characterization, provided by the various measures of central tendency of the posterior distribution. In particular, the mean, median, or mode of the posterior distribution might be chosen to communicate a point estimate for θ. In the case of the beta posterior distribution in Equation 6.6, the posterior mean is $6/17 = 0.353$ (Equation 4.59a), the median (which could be found through numerical integration, or tables such as those in Winkler, 1972b) is 0.347, and the posterior mode (value of p maximizing the posterior distribution) is 0.333.

The posterior mode can be an especially attractive point estimate because of its relationship to the maximum likelihood estimate for θ (Section 4.6). The influence of the prior distribution on the posterior distribution becomes quite small for problems where large amounts of data are available, so that the posterior distribution becomes nearly proportional to the likelihood alone. In that case the posterior mode is nearly the same as the value of θ maximizing the likelihood. In the case of a uniform prior distribution the posterior distribution is exactly proportional to the likelihood (Equation 6.2, with $f(\theta) = 1$), so that the posterior mode in Example 6.1 is exactly the maximum likelihood estimate for the binomial probability: $\hat{p} = 5/15 = 0.333$, having observed $x = 5$ successes in $N = 15$ trials.

Of course summarizing the posterior distribution using probabilities is more informative than is a single-number expression of central tendency. Most commonly this is done using a *central credible interval* (CCI), which will span a range for θ corresponding (in probability) to the middle portion of the posterior distribution. For example, the 95% CCI for the beta posterior distribution with $\alpha = 6$ and $\beta = 11$ in Figure 6.1 is [0.152, 0.587], as indicated by the double-headed arrow. These endpoints are calculated as the $q_{.025}$ and the $q_{.975}$ quantiles of the posterior distribution. The interpretation of this interval is that there is a 0.95 probability that θ lies within it. For many people this is a more natural

inferential interpretation than the repeated-sampling concept associated with the $(1-\alpha)\times 100\%$ frequentist confidence interval (Section 5.1.7), and indeed many people incorrectly ascribe this meaning of the Bayesian credible interval to frequentist confidence intervals (e.g., Ambaum, 2010).

An alternative, although generally more computationally difficult, credible interval is the *highest posterior density* (HPD) interval. The HPD interval also spans a specified amount of probability, but is defined with respect to the largest possible corresponding values of the posterior distribution. Imagine a horizontal line intersecting the posterior density, and thus defining an interval. The HPD interval corresponding to a given probability is defined by the two points of intersection of that horizontal line with the posterior density, for which the given probability is just encompassed by the interval. An HPD interval can thus be viewed as a probabilistic extension of the posterior mode. For a symmetric posterior distribution the HPD interval will coincide with the simpler CCI. For a skewed posterior distribution (such as in Figure 6.1), the HPD interval will be somewhat shifted and compressed relative to the CCI.

In some settings the probability that θ may be above or below some physically meaningful level could be of interest. In such cases the most informative summary of the posterior distribution might simply be a computation of the probability that θ is above or below the threshold.

6.2.3. The Role of the Prior Distribution

The prior distribution $f(\theta)$ quantitatively characterizes the analyst's uncertainty or degree of belief about possible values of the parameter θ, before new data become available. It is a potentially controversial element of Bayesian inference, because different people can reasonably differ in their judgments, and thus can reasonably hold prior beliefs that are different from each other. If the available data are relatively few, then different priors may lead to quite different posterior distributions, and thus to quite different inferences about θ. On the other hand, in data-rich settings the influence of the prior distribution is relatively much less important, so that inferences derived from most reasonable priors will be very similar to each other.

Accurately quantifying prior beliefs may be a difficult task, depending on the circumstances and the experience of the analyst. It is not necessary for a prior distribution to be of a known or familiar parametric form, for example, one of the distributions presented in Chapter 4. One approach to assessing subjective probability is through the use of hypothetical betting or "lottery" games in order to refine one's judgments about probabilities for discrete events (Section 7.10.4) or quantiles of continuous probability distributions (Section 7.10.5). In the continuous case, the subjectively elicited quantiles may provide a basis for constructing a continuous mathematical function representing the relative prior beliefs. Because of the equivalence between Equations 6.1 and 6.2 such functions need not necessarily be proper probability densities, although depending on the form of the function the normalizing constant in the denominator of Equation 6.1 may be difficult to compute.

Sometimes it is both conceptually and mathematically convenient to adopt a known parametric form for the prior distribution, and then to subjectively elicit its parameters (i.e., the prior hyperparameters) based on the properties of the chosen distributional form. For example, if one is able to form a judgment regarding the mean or median of one's prior distribution, this can provide a useful constraint on the prior hyperparameters. Certain parametric forms that are compatible with a particular data-generating model (i.e., the likelihood appropriate to a given problem) may greatly simplify the subsequent calculations, as discussed in Section 6.3, although a mathematically convenient prior that is a poor approximation to one's subjective judgments should not be chosen.

Another important aspect of the prior distribution relates to specification of zero probability for some of the mathematically allowable values of θ. This quite strong condition will usually not be justified, because any range of values for θ assigned zero probability by the prior cannot have nonzero probability in the posterior distribution, regardless of the strength of the evidence provided by the data. This point can be appreciated by examining Equations 6.1 or 6.2: any values of θ for which $f(\theta) = 0$ will necessarily yield $f(\theta|x) = 0$, for all possible data x. Any values of θ that cannot absolutely be ruled out by prior information (e.g., by constraints implied by the underlying physics, such as negative Kelvin temperatures) should be assigned nonzero (although possibly extremely small) probability in the prior distribution.

In situations where there is very little prior information with which to judge relative plausibility for different values of θ, it is natural to choose a prior distribution that does not favor particular values over others to an appreciable degree; that is, a prior distribution expressing as nearly as possible a state of ignorance. Such prior distributions are called *diffuse priors, vague priors, flat priors,* or *noninformative priors*. The prior distribution $f(p) = 1$ in Example 6.1 is an example of such a prior distribution.

Diffuse prior distributions are sometimes seen as being more objective, and therefore less controversial than priors expressing specific subjective judgments. In part this conclusion derives from the fact that a diffuse prior influences the posterior distribution to a minimum degree, by giving maximum weight in Bayes' Theorem to the (data-controlled) likelihood. In general the evidence provided in the data will overwhelm a diffuse prior unless the data sample is fairly small. As has already been noted, Bayesian inference with a diffuse prior will then usually be similar to inferences based on maximum likelihood.

When the parameter θ of interest is not bounded, either above or below or both, it may be difficult to construct a diffuse prior that is consistent with an analyst's subjective judgments. For example, if the parameter of interest is a Gaussian mean, its possible values include the entire real line. One possibility for a diffuse prior in this case could be a Gaussian distribution with zero mean and a very large but finite variance. This prior distribution is nearly flat, but still slightly favors values for the mean near zero. Alternatively, it might be useful to use an *improper prior*, having the property $\int_\theta f(\theta)\, d\theta \neq 1$, such as $f(\theta) = \text{constant}$ for $-\infty < \theta < \infty$. Surprisingly, improper priors do not necessarily lead to nonsense inferences, because of the equivalence of Equations 6.1 and 6.2. In particular, an improper prior is permissible if the integral in the denominator of Equation 6.1 yields a finite nonzero value, so that the resulting posterior distribution is a proper probability distribution, with $\int_\theta f(\theta|x)\, d\theta = 1$.

6.2.4. The Predictive Distribution

The ultimate goal of some inferential analyses will be to gain insight about future, yet-unobserved values of the data x^+, which in turn will be informed by the quantification of uncertainty regarding the parameter (s) θ. That is, we may wish to make probability forecasts for future data values that account both for the way their generating process varies for different values of θ, and for the relative plausibility of different values of θ provided by the posterior distribution.

The *predictive distribution* is a probability density function for future data that is derived from a combination of the parametric data-generating process and the posterior distribution for θ,

$$f(x^+) = \int_\theta f(x^+|\,\theta) f(\theta|\,x) d\theta. \tag{6.7}$$

Here x^+ denotes the future, yet-unobserved data, and x represents the data that has already been used in Bayes' Theorem to produce the current posterior distribution $f(\theta|x)$. Since Equation 6.7 expresses the

unconditional PDF (if x is continuous) or probability distribution function (if x is discrete), $f(x|\theta)$ quantifies the data-generating process. It is the PDF (or probability distribution function) for the data given a particular value of θ, not the likelihood for θ given a fixed data sample x, although as before the two are notationally the same. The posterior PDF $f(\theta|x)$ quantifies uncertainty about θ according to the most recently available probability updating, and accordingly Equation 6.7 is sometimes called the *posterior predictive distribution*. If Equation 6.7 is to be applied before observing any data, $f(\theta|x)$ will be the prior distribution, in which case Equation 6.7 will be notationally equivalent to the denominator in Equation 6.1.

Equation 6.7 yields an unconditional PDF for future data x^+ that accounts both for the uncertainty about θ and uncertainty about x for each possible value of θ. It is in effect a weighted average of the PDFs $f(x^+|\theta)$ for all possible values of θ, where the weights are provided by posterior distribution. If θ could somehow be known with certainty, then $f(\theta|x)$ would put probability 1 on that value, and Equation 6.7 would simply be equal to the data-generating PDF $f(x^+|\theta)$ evaluated at that θ. However, Equation 6.7 explicitly accounts for the effects of uncertainty about θ, yielding increased uncertainty about future values of x consistent with the uncertainty about θ.

6.3. CONJUGATE DISTRIBUTIONS

6.3.1. Definition of Conjugate Distributions

An appropriate mathematical form for the likelihood in Equations 6.1 and 6.2, which characterizes the data-generating process, is often clearly dictated by the nature of the problem at hand. However, the form of the prior distribution is rarely so well defined, depending as it does on the judgment of the analyst. In this general case, where the form of the prior distribution is not constrained by the form of the likelihood, evaluation of Equations 6.1 and 6.7 may require numerical integration or other computationally intensive methods, and the difficulty is compounded if the probability updating must be computed iteratively rather than only once.

For certain mathematical forms of the likelihood, however, the computations of Bayes' Theorem can be greatly simplified if choice of a *conjugate distribution* for the prior to be used with that likelihood can be justified. A prior distribution that is conjugate to a particular likelihood is a parametric distribution that is similar mathematically to that likelihood, in a way that yields a posterior distribution that has the same parametric form as the prior distribution. Use of a conjugate distribution that is compatible with a given data-generating process greatly simplifies the computations associated with Bayesian inference by allowing closed-form expressions for the posterior PDF. In addition, simple relationships between the hyperparameters of the prior and posterior distributions can provide insights into the relative importance to the posterior distribution of the prior distribution and the available data. Use of conjugate distributions also facilitates iterative updating of Bayes' Theorem, since the previous posterior distribution, which becomes the new prior distribution when additional data become available, is of the same conjugate parametric form.

Choosing to work with a conjugate prior is convenient, but represents a strong constraint on how the analyst's prior beliefs can be expressed. When the parametric form of the conjugate distribution is very flexible there can be broad scope to approximate the analyst's actual prior beliefs, but an adequate representation is not guaranteed. On the other hand, representation of subjective beliefs using any mathematically explicit PDF will nearly always be an approximation, and the degree to which a nonconjugate prior might be a better approximation may be balanced against the advantages provided by conjugate distributions.

The following sections outline Bayesian inference using conjugate distributions, for three simple but important data-generating processes: the binomial, Poisson, and Gaussian distributions.

6.3.2. Binomial Data-Generating Process

When the data of interest consist of the numbers of "successes" x obtained from N independent and identically distributed Bernoulli trials, their probability distribution will be binomial (Equation 4.1). In this setting the inferential question typically pertains to the value of the success probability (p in Equation 4.1). The appropriate likelihood is then given by the first equality in Equation 6.3, which is notationally identical to Equation 4.1, but is a function of the success probability p given a fixed number of successes x in N independent realizations.

The conjugate prior distribution for the binomial data-generating process is the beta distribution (Equation 4.58). According to Equation 6.2, we can ignore the scaling constants $\binom{N}{x}$ and $\Gamma(\alpha+\beta)/[\Gamma(\alpha)\Gamma(\beta)]$ in Equations 4.1 and 4.58, respectively, so that Bayes' Theorem for the binomial data-generating process and a beta prior distribution becomes

$$f(p|x) \propto p^x(1-p)^{N-x}\, p^{\alpha-1}p^{\beta-1} = p^{x+\alpha-1}(1-p)^{N-x+\beta-1}. \tag{6.8}$$

Here p is the Bernoulli success probability about which inferences are being computed, and α and β are the hyperparameters of the beta prior distribution. Because of the similarity in mathematical form (apart from the terms not involving p) between the binomial likelihood and the beta prior distribution, their product simplifies to the final equality in Equation 6.8. This simplification shows that the posterior distribution for the success probability, $f(p|x)$, is also a beta distribution, with hyperparameters

$$\alpha' = x+\alpha \tag{6.9a}$$

and

$$\beta' = N-x+\beta. \tag{6.9b}$$

Adopting the conjugate prior has allowed evaluation of Equation 6.1 using just these two simple relationships, rather than requiring a potentially difficult integration or some other computationally demanding procedure. Including the scaling constant to ensure that the posterior PDF integrates to 1,

$$f(p|x) = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(x+\alpha)\,\Gamma(N-x+\beta)} p^{x+\alpha-1}(1-p)^{N-x+\beta-1}. \tag{6.10}$$

The relationship between the hyperparameters of the prior beta distribution, α and β, to the hyperparameters of the posterior beta distribution, Equations 6.9, illustrates a more general attribute of Bayesian inference. As more data accumulate, the posterior distribution depends progressively less on whatever choice has been made for the prior distribution (assuming that possible ranges of θ have not been assigned zero prior probability). In the present case of binomial inference with a conjugate prior, $x \gg \alpha$ and $N-x \gg \beta$ if a sufficiently large amount of data can be collected. Therefore the posterior density approaches the binomial likelihood (again apart from the scaling constants), since in that case $x \approx x+\alpha-1$ and $N-x \approx N-x+\beta-1$.

Although not mentioned at the time, Example 6.1 was computed using a conjugate prior distribution, because the uniform distribution $f(p)=1$ is a special case of the beta distribution with hyperparameters

$\alpha = \beta = 1$, and this is exactly the reason that Equations 6.4 and 6.6 are also beta distributions. Equation 6.10 also illustrates clearly why the posterior distribution in Equation 6.6 was achieved regardless of whether Bayes' Theorem was applied individually for each of the two data batches as was done in Example 6.1, or only once after having observed $x = 5$ successes in the overall total of $N = 15$ realizations. In the latter case, the hyperparameters of the posterior beta distribution are also $x + \alpha = 5 + 1 = 6$ and $N - x + \beta = 15 - 5 + 1 = 11$. Since $\alpha = \beta = 1$ yields $f(p) = 1$ for the prior distribution, the posterior distributions in Example 6.1 are exactly proportional to the corresponding binomial likelihoods, which is why the posterior modes are equal to the corresponding maximum likelihood estimates for p. For beta distributions where $\alpha > 1$ and $\beta > 1$, the mode occurs at $(\alpha - 1)/(\alpha + \beta - 2)$.

The influence of uncertainty about the binomial success probability on the probability distribution for future numbers of successes, x^+, among N^+ future realizations, is quantified through the predictive distribution. In the setting of binomial likelihood and a conjugate beta prior distribution, Equation 6.7 is evaluated after substituting the binomial probability distribution function (Equation 4.1) with success probability p, for $f(x^+|\theta)$, and the posterior beta distribution from Equation 6.10 for $f(\theta|x)$. The result is the discrete probability distribution function

$$\Pr\{X^+ = x^+\} = \binom{N^+}{x^+} \left[\frac{\Gamma(N+\alpha+\beta)}{\Gamma(x+\alpha)\,\Gamma(N-x+\beta)}\right] \frac{\Gamma(x^+ + x + \alpha)\Gamma(N^+ + N - x^+ - x + \beta)}{\Gamma(N^+ + N + \alpha + \beta)} \tag{6.11a}$$

$$= \binom{N^+}{x^+} \left[\frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\,\Gamma(\beta')}\right] \frac{\Gamma(x^+ + \alpha')\Gamma(N^+ - x^+ + \beta')}{\Gamma(N^+ + \alpha' + \beta')}, \tag{6.11b}$$

known as the *beta-binomial*, or *Polya distribution*. This function distributes probability among the possible integer outcomes $0 \le x^+ \le N^+$. In Equation 6.11a, α and β are the hyperparameters for the prior beta distribution pertaining to p, and x indicates the number of successes in the N data realizations used to update that prior to the posterior distribution in Equation 6.10. The beta-binomial distribution in Equation 6.11b can be thought of as the probability distribution function for a binomial variable, when the success probability p is drawn randomly for each realization from the posterior beta distribution with hyperparameters α' and β'. The mean and variance for the beta-binomial distribution are

$$\mu = \frac{N^+(x+\alpha)}{N+\alpha+\beta} \tag{6.12a}$$

$$= \frac{N^+\alpha'}{\alpha'+\beta'} \tag{6.12b}$$

and

$$\sigma^2 = \frac{N^+(x+\alpha)(N-x+\beta)(N^+ + N + \alpha + \beta)}{(N+\alpha+\beta)^2(N+\alpha+\beta+1)} \tag{6.13a}$$

$$= \frac{N^+\alpha'\beta'(N^+ + \alpha' + \beta')}{(\alpha'+\beta')^2(\alpha'+\beta'+1)}. \tag{6.13b}$$

Example 6.2 Bayesian Reanalysis of Example 5.1

Example 5.1 considered a hypothetical situation in which the claim that the climatological probability of a cloudless day in winter is 6/7 was examined, after observing $x = 15$ cloudless days on $N = 25$

independent occasions. Analysis of this situation in a Bayesian framework is straightforward if the analyst's prior uncertainty about the winter sunshine climatology at this location can be characterized with a beta distribution. Because beta distributions are able to represent a wide variety of shapes on the unit interval, they can often provide good approximations to an individual's subjective degree of belief about the true value of a probability, such as the binomial success probability, p.

Consider the effects of two possible prior distributions for this probability. First, someone with little or no knowledge of the context of this analysis might reasonably adopt the diffuse uniform prior distribution, equivalent to the beta distribution with $\alpha = \beta = 1$. Someone who is more sophisticated about the nature of advertising claims might use this prior knowledge to form the judgment that there might only be a 5% chance of this binomial p being above the claimed 6/7 value. If in addition this second individual thought that values of p above and below 0.5 were equally plausible (i.e., thinking the median of their prior distribution is 0.5), these two conditions together would fully determine a beta prior with $\alpha = \beta = 4$.

Because both of these two priors are beta distributions, it is straightforward to use Equation 6.10 to compute the posterior distributions after having observed $x = 15$ successes in $N = 25$ independent Bernoulli trials. Because the beta distribution is conjugate to the binomial likelihood, both of these posterior distributions are also beta distributions. The uniform prior is updated by Equation 6.10 to the beta distribution with $\alpha' = 16$ and $\beta' = 11$, and the $\alpha = \beta = 4$ prior distribution is updated by these same data to the posterior beta distribution with $\alpha' = 19$ and $\beta' = 14$.

These two posterior distributions and their corresponding priors are shown in Figure 6.2a and b. Although the two prior distributions are quite different from each other, even the modest amount of data used to update them has been sufficient for the two posterior distributions to be quite similar. For the posterior distribution in Figure 6.2a, the mode $[= (16-1)/(16+11-2) = 15/25 = 0.600]$ is exactly the maximum likelihood estimate for p because the prior $f(p) = 1$, so that the posterior is exactly proportional to the likelihood. In Figure 6.2b the posterior mode is 0.581, which is different from but still similar to the posterior mode in Figure 6.2a. Although the two posterior distributions in Figure 6.2 are similar, the sharper prior information in Figure 6.2b leads to a somewhat more concentrated (lower-variance) posterior distribution. This difference is reflected by the corresponding 95% CCIs, which are [0.406, 0.776] in Figure 6.2a and [0.406, 0.736] in Figure 6.2b. The claimed probability of $p = 6/7$ is quite implausible according to both of these posterior analyses, with $\Pr\{p \geq 6/7\} = 0.00048$

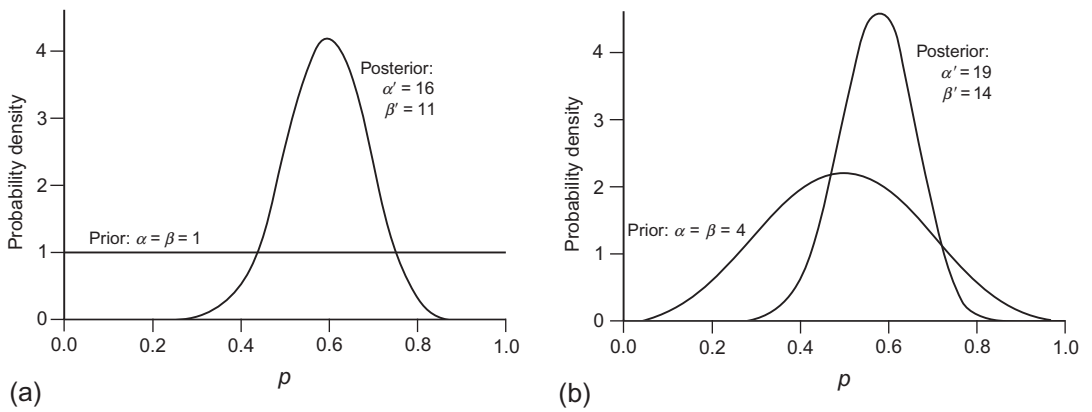

FIGURE 6.2 Comparison of posterior beta densities after having observed $x = 15$ successes in $N = 25$ Bernoulli trials, when (a) the prior beta density is uniform ($\alpha = \beta = 1$), and (b) the prior beta density has parameters $\alpha = \beta = 4$.

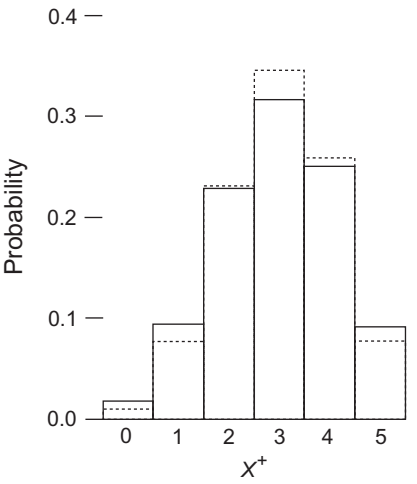

FIGURE 6.3 Beta-binomial predictive distribution with $\alpha' = 16$ and $\beta' = 11$ for the number of cloudless days X^+ in the next $N^+ = 5$ independent observations (solid histogram), compared to binomial probabilities obtained with $p = 0.6$ (dashed).

according to the posterior distribution in Figure 6.2a, and $\Pr\{p \geq 6/7\} = 0.000054$ in Figure 6.2b. Both results are generally consistent with the conclusion reached in Example 5.1.

In addition to inferences regarding the parameter p of the binomial data-generating process, in many situations it might also be of interest to make inferences about the probability distribution for future data values, which are quantified by the predictive distribution. For inferences about the binomial data-generating process that have been computed using conjugate beta distributions, the predictive distributions are beta-binomial distributions, Equation 6.11.

Suppose we are interested in the possible numbers of cloudless days, X^+, in the next $N^+ = 5$ independent observations of the sky condition at this desert resort, according to the posterior distribution in Figure 6.2a, with $\alpha' = 16$ and $\beta' = 11$. This will be a discrete distribution with $N^+ + 1 = 6$ possible outcomes, as indicated by the solid histogram bars in Figure 6.3. Not surprisingly the most likely outcome is $X^+ = 3$ cloudless days out of $N^+ = 5$. However, there are nonzero probabilities for the other 5 outcomes also, and the distribution of probability among the outcomes reflects both sampling variability deriving from the 5 Bernoulli trials, as well as uncertainty about the actual value of the Bernoulli success probability, p, that is quantified by the posterior distribution. The effect of this latter source of uncertainty can be appreciated by comparing the dotted histogram in Figure 6.3, which portrays the probabilities from the binomial distribution with $p = 0.6$ and $N = 5$. This binomial distribution would be the predictive distribution if it could be known with certainty that the success probability is 0.6, but uncertainty about p leads to additional uncertainty about X^+, so that the beta-binomial predictive distribution in Figure 6.3 allocates less probability to the middle values of X^+ and more probability to the extreme values. ◇

Both the geometric distribution (Equation 4.5) and the negative binomial (Equation 4.6) distribution are closely related to the binomial data-generating process, since all three pertain to outcomes of independent Bernoulli trials. Looking more closely at these two probability distribution functions, it can be seen that the corresponding likelihood functions (again, apart from scaling constants not depending on the success probability p) are notationally analogous to the PDF for the beta distribution (again, apart from the scaling constants involving the gamma functions). As would be suggested by this similarity, beta distributions provide conjugate priors for these data-generating processes as well, allowing convenient Bayesian inference in these settings. Epstein (1985) provides the predictive distribution for the

Pascal (negative binomial distribution with integer parameter) data-generating process when a beta prior distribution is used, called the *beta-Pascal distribution*.

6.3.3. Poisson Data-Generating Process

The Poisson data-generating process (Section 4.2.4) is also amenable to simplification of Bayesian inference using conjugate prior distributions. In this case the parameter that is the subject of inference is the Poisson mean, μ, which specifies the average rate of event occurrences per unit interval (usually, a time interval). Rewriting the form of Equation 4.12 as a function of μ, and omitting the denominator that does not depend on it, the Poisson likelihood is proportional to

$$f(x|\,\mu) \propto \mu^x \exp[-\mu]. \tag{6.14}$$

This likelihood is mathematically similar to the PDF of the gamma distribution (Equation 4.45, with the mean μ as the random variable) which, again excluding factors not depending on μ, is proportional to

$$f(\mu) \propto \mu^{\alpha-1} \exp[-\mu/\beta]. \tag{6.15}$$

The two factors on the right-hand sides of Equations 6.14 and 6.15 combine when multiplied together in Equation 6.2, so that the gamma distribution is conjugate to the Poisson likelihood. Therefore when a gamma prior distribution for μ with hyperparameters α and β can be reasonably assumed (i.e., is consistent with a particular analyst's judgments, to good approximation), the resulting posterior distribution will also be gamma, and proportional to

$$f(\mu|\,x) \propto f(x|\,\mu) f(\mu) \propto \mu^x \exp[-\mu]\,\mu^{\alpha-1} \exp[-\mu/\beta] = \mu^{x+\alpha-1} \exp[-(1+1/\beta)\mu]. \tag{6.16}$$

The likelihood in Equation 6.14 pertains to the number of observed events, x, in a single unit time interval. Often the available data will consist of the total number of event counts over multiple (say, n) independent time intervals. In such cases the likelihood for the total number of events during the n time units will be the product of n likelihoods of the form of Equation 6.14. Denoting now the total number of events in these n time intervals as x, that Poisson likelihood is proportional to

$$f(x|\,\mu) \propto \mu^x \exp[-n\mu], \tag{6.17}$$

which when combined with a gamma prior distribution for μ (Equation 6.15) yields the posterior distribution

$$f(\mu|\,x) \propto f(x|\,\mu) f(\mu) \propto \mu^x \exp[-n\mu]\mu^{\alpha-1} \exp[-\mu/\beta] = \mu^{x+\alpha-1} \exp[-(n+1/\beta)\mu]. \tag{6.18}$$

Comparing the final expression in Equation 6.18 with Equation 4.45 it is clear that this posterior distribution is also a gamma distribution, with hyperparameters

$$\alpha' = \alpha + x \tag{6.19a}$$

and, since $1/\beta' = 1/\beta + n$,

$$\beta' = \frac{\beta}{1+n\beta}. \tag{6.19b}$$

The resulting posterior gamma PDF can therefore be expressed either in terms of the prior hyperparameters and the data,

$$f(\mu|x) = \frac{\left[\left(\frac{1}{\beta}+n\right)\mu\right]^{\alpha+x-1} \exp\left[-\left(\frac{1}{\beta}+n\right)\mu\right]}{\left(\frac{\beta}{1+n\beta}\right)\Gamma(\alpha+x)}, \tag{6.20a}$$

or in terms of the posterior hyperparameters in Equation 6.19,

$$f(\mu|x) = \frac{(\mu/\beta')^{\alpha'-1}\exp(-\mu/\beta')}{\beta'\,\Gamma(\alpha')}. \tag{6.20b}$$

As could also be seen in Equation 6.9 for the conjugate hyperparameters for the binomial data-generating process, Equation 6.19 shows that progressively larger amounts of data yield posterior gamma distributions that are less influenced by the prior hyperparameters α and β. In particular, as x and n both become large, $\alpha' \approx x$ and $\beta' \approx 1/n$. The dependence on the prior distribution is further lessened when the prior is diffuse. One possibility for a diffuse prior gamma distribution is $f(\mu) \propto 1/\mu$, which is uniform in $\ln(\mu)$. This is an improper prior distribution, but corresponds formally to the prior hyperparameters $\alpha = 1/\beta = 0$, so that Equation 6.19b yields $\alpha' = x$ and $\beta' = 1/n$, exactly, for the resulting posterior hyperparameters.

The predictive distribution, Equation 6.7, for (the discrete) numbers of future Poisson events $x^+ = 0, 1, 2, \ldots$ in a given future unit interval, is the negative binomial distribution

$$\Pr\{X^+ = x^+\} = \frac{\Gamma(x^+ + \alpha')}{\Gamma(\alpha')\,x^+!}\left(\frac{1}{1+\beta'}\right)^{\alpha'}\left(\frac{\beta'}{1+\beta'}\right)^{x^+}. \tag{6.21}$$

This is of the same form as Equation 4.6, where the probability p has been parameterized in Equation 6.21 as $1/(1+\beta')$. This result for the predictive distribution points out another interpretation for the negative binomial distribution, namely, that it describes a Poisson distribution with a random rate parameter μ, that is drawn anew for each time interval from the gamma distribution with parameters α' and β'. That is, the predictive distribution in Equation 6.21 accounts both for the interval-to-interval variability in the number of Poisson events for a particular value of the rate parameter μ, and for uncertainty about μ that is quantified by its gamma posterior distribution.

Example 6.3 Poisson Mean for U.S. Landfalling Hurricanes

It was noted in Example 4.4 that the Poisson is a natural distribution for characterizing the data-generating process for annual numbers of hurricanes making landfall in the United States. However, sample estimates of the Poisson rate μ must be based on the available data for annual U.S. hurricane counts, and are therefore subject to some uncertainty. These data are available from 1851 onward, but estimation of the Poisson rate for this phenomenon is complicated by the fact that the earlier data are generally believed to be less reliable.

One approach to dealing with the uneven reliability of the historical annual hurricane count data might be to focus only on the more recent years and ignore the older values. Elsner and Bossak (2001) suggested an alternative approach that makes use of the earlier data without assuming that it is of the same quality as the later data. Their approach was to use the earlier (1851–99) and less reliable data to estimate a prior distribution for the Poisson mean, and then to revise this prior distribution in light of the remaining (1900–2000) data, using Bayes' Theorem.

To specify their prior distribution, Elsner and Bossak (2001) bootstrapped (Section 5.3.5) the 1851–99 annual U.S. landfalling hurricane counts to estimate the sampling distribution for the mean annual number,

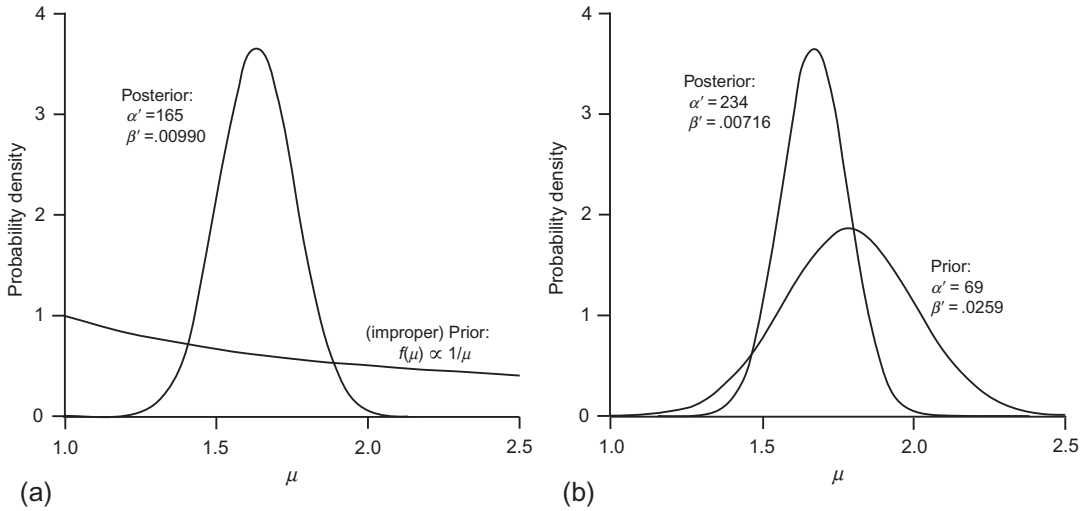

FIGURE 6.4 Posterior gamma PDFs for the Poisson mean characterizing annual numbers of U.S. landfalling hurricanes, resulting from updating (a) the diffuse, improper prior proportional to $1/\mu$, and (b), a gamma prior derived from bootstrapping hurricane landfall counts from the years 1851–99.

μ. The 5th and 95th percentiles of this estimated sampling distribution are 1.45 and 2.16 hurricanes per year, respectively, which quantiles are consistent with a gamma prior distribution with $\alpha = 69$ and $\beta = 0.0259$. The mean of this distribution (Table 4.5) is $\alpha\beta = (69)(0.0259) = 1.79$ hurricanes per year, which agrees well with the sample mean of 1.76 hurricanes per year for the years 1851–99.

For the $n = 101$ years 1900–2000, there were $x = 165$ U.S. landfalling hurricanes. Substituting these values into Equation 6.19, together with the prior hyperparameters $\alpha = 69$ and $\beta = 0.0259$, yields the gamma posterior hyperparameters $\alpha' = 234$ and $\beta' = 0.00716$. Alternatively, adopting the diffuse prior $\alpha = 1/\beta = 0$ leads to the gamma posterior distribution with $\alpha' = 165$ and $\beta' = 0.00990$. Figure 6.4 compares these prior-posterior pairs. Because both have large shape parameter α', each is closely approximated by a Gaussian distribution. The Elsner and Bossak (2001) posterior distribution in Figure 6.4b has a posterior mode of 1.668 (the mode of the gamma distribution, for $\alpha > 1$, is $\beta(\alpha - 1)$), and its 95% CCI is [1.46, 1.89]. The posterior distribution computed from the diffuse prior (Figure 6.4a) is similar but somewhat less sharp, having its mode at $(0.00990)(164–1) = 1.614$, with a 95% CCI of [1.38, 1.88]. The additional information in the nondiffuse prior distribution in Figure 6.4b has resulted in a lower-variance posterior distribution, exhibiting somewhat less uncertainty about the Poisson rate.

The probability distribution for numbers of U.S. landfalling hurricanes in some future year, accounting for both year-to-year differences in numbers of realized Poisson events, and uncertainty about their mean rate characterized by a gamma posterior distribution, is the negative binomial predictive distribution in Equation 6.21. Direct evaluation of Equation 6.21 for the present example is problematic, because the large arguments in the gamma functions will lead to numerical overflows. However, this problem can be circumvented by first computing the logarithms of the probabilities for each of the x^+ of interest, using series representations for the logarithm of the gamma function (e.g., Abramowitz and Stegun 1984, Press et al. 1986).

Figure 6.5 compares the negative binomial predictive distribution (solid histogram), computed using the posterior distribution in Figure 6.4b, to the Poisson distribution (dashed) with mean $\mu = 165/101 = 1.634$ (the annual average number of U.S. hurricane landfalls, 1900–2000). The two

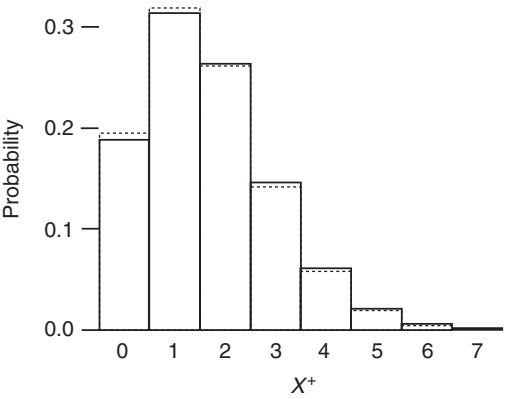

FIGURE 6.5 Negative binomial predictive distribution with $\alpha' = 234$ and $\beta' = 0.00716$ for the number of U.S. landfalling hurricanes (*solid histogram*), compared to Poisson probabilities obtained with $\mu = 165/101 = 1.634$ (*dashed*).

distributions are quite close, reflecting the rather compact character of the posterior distribution in Figure 6.4b, although the negative binomial predictive distribution has a slightly larger variance ($\sigma^2 = 1.687$, cf. Table 4.3) than the Poisson distribution ($\sigma^2 = \mu = 1.634$), which is reflected by the longer right tail. ◇

6.3.4. Gaussian Data-Generating Process

Bayesian inference for the mean μ of a Gaussian (Equation 4.24) data-generating process is also amenable to analytic treatment using conjugate prior and posterior distributions. The general case, where both the mean μ and variance σ^2 of the generating process are unknown, becomes quite complicated because the joint posterior distribution of the two parameters must be considered, even if their univariate prior distributions $f(\mu)$ and $f(\sigma^2)$ can reasonably be regarded as independent. Treatments of that case can be found in Epstein (1985) and Lee (1997), for example.

The more restricted case, where inferences about a Gaussian μ, assuming that the variance of the data-generating process is known, is much more straightforward. Instances where this assumption may be justified include analysis of data produced by an instrument whose measurement precision is well known, or in large-sample settings where the sample variance is known to estimate the variance of the generating process very closely.

An interesting aspect of Bayesian inference for the mean of a Gaussian generating process, assuming known variance, is that the conjugate prior and posterior distributions are also Gaussian. Furthermore, when the posterior distribution is Gaussian, then the predictive distribution is Gaussian as well. This situation is computationally convenient, but notationally confusing because four sets of means and variances must be distinguished. In the following, the symbol μ will be used for the mean of the data-generating process, about which inferences are to be made. The known variance of the data-generating process will be denoted as σ_*^2. The hyperparameters of the prior Gaussian distribution will be denoted as μ_h and σ_h^2, respectively, and will be distinguished from the posterior hyperparameters μ_h' and $\sigma_h^{2\prime}$. The parameters of the Gaussian predictive distribution will be represented by μ_+ and σ_+^2.

Using this notation the prior distribution is proportional to

$$f(\mu) \propto \frac{1}{\sigma_h} \exp\left[-\frac{(\mu-\mu_h)^2}{2\sigma_h^2}\right], \tag{6.22}$$

and the likelihood, given a data sample of n independent values x_i from the data-generating process, is proportional to

$$f(x|\,\mu) \propto \prod_{i=1}^{n} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma_*^2}\right]. \tag{6.23a}$$

However, the sample mean carries all the relevant information in the data pertaining to μ (the sample mean is said to be sufficient for μ), so that the likelihood can be expressed more compactly as

$$f(\overline{x}|\,\mu) \propto \exp\left[-\frac{n(\overline{x}-\mu)^2}{2\sigma_*^2}\right], \tag{6.23b}$$

because the distribution for a sample mean of n data values from a Gaussian distribution with parameters μ and σ_*^2 is itself Gaussian, with mean μ and variance σ_*^2/n. Combining Equations 6.22 and 6.23b using Bayes' Theorem leads to the Gaussian posterior distribution for μ,

$$f(\mu|\,\overline{x}) = \frac{1}{\sqrt{2\pi}\sigma_{h'}} \exp\left[-\frac{(\mu-\mu_h{}')^2}{2\sigma_h^2{}'}\right], \tag{6.24}$$

where the posterior hyperparameters are

$$\mu_h{}' = \frac{\mu_h/\sigma_h^2 + n\overline{x}/\sigma_*^2}{1/\sigma_h^2 + n/\sigma_*^2} \tag{6.25a}$$

and

$$\sigma_h^2{}' = \left(\frac{1}{\sigma_h^2} + \frac{n}{\sigma_*^2}\right)^{-1}. \tag{6.25b}$$

That is, the posterior mean is a weighted average of the prior mean and the sample mean, with progressively greater weight given to the sample mean as n increases. The reciprocal of the posterior variance is the sum of the reciprocals of the prior variance and the (known) data-generating variance, so that the posterior variance is necessarily smaller than both the prior variance and the data-generating variance, and decreases as n increases. Only the sample mean, and not the sample variance, appears in Equation 6.25 for the posterior parameters because of the assumption that σ_*^2 is known, so that no amount of additional data can improve our knowledge about it.

For analyses where diffuse prior distributions are appropriate, the most common approach when using a Gaussian prior distribution is to specify an extremely large prior variance, so that the prior distribution is nearly uniform over a large portion of the real line around the prior mean. In the limit of $\sigma_h^2 \to \infty$, the resulting diffuse prior distribution is uniform on the real line, and therefore improper. However, this choice yields $1/\sigma_h^2 = 0$ in Equation 6.25, so the posterior distribution is proportional to the likelihood, with $\mu_h{}' = \overline{x}$ and $\sigma_h^{2\prime} = \sigma_*^2/n$.

Uncertainty about future data values x^+ from the Gaussian data-generating process results from the combination of sampling variability from the data-generating process itself in combination with uncertainty about μ that is expressed by the posterior distribution. These two contributions are quantified by the predictive distribution, which is also Gaussian, with mean

$$\mu_+ = \mu_h' \tag{6.26a}$$

and variance

$$\sigma_+^2 = \sigma_*^2 + \sigma_h^{2\prime}. \tag{6.26b}$$

Example 6.4 Bayesian Inference for Wind power Suitability

Before wind turbines for generation of electricity are purchased and installed at a location, an evaluation of the suitability of the local climate for wind power generation is prudent. A quantity of interest in this evaluation may be the average *wind power density* at 50 m height. Suppose a wind farm will be economically viable if the average annual wind power density is at least 400 W/m^2. Ideally a long climatological record of wind speeds would be very helpful in evaluating the suitability of a candidate site, but practically it may be possible to set up an anemometer to make wind measurements at a potential wind power site for only a year or two before the decision is made. How might such measurements be used to evaluate the wind power suitability?

The wind power density depends on the cube of wind speed, the distribution of which is usually positively skewed. However, when averaged over a long time period such as a year, the Central Limit Theorem suggests that the distribution of the annual average will be at least approximately Gaussian. Suppose previous experience with other wind farms is that the year-to-year variability in the annually averaged wind power density can be characterized by a standard deviation of 50 W/m^2. These conditions suggest a Gaussian data-generating process for the annual average wind power density at a location, with unknown mean μ and known standard deviation $\sigma_* = 50$ W/m^2.

Someone contemplating construction of a new wind power site will have some prior belief regarding possible values for μ. Suppose this person's prior distribution for μ is Gaussian, with mean $\mu_h = 550$ W/m^2. If in addition this person's judgment is that there is only a 5% chance that μ will be smaller than 200 W/m^2, the implied prior standard deviation is $\sigma_h = 212$ W/m^2.

Suppose now that it is possible to collect $n = 2$ years of wind data before deciding whether or not to begin construction, and that the average wind power densities for these 2 years are 420 and 480 W/m^2. These are certainly consistent with the degree of interannual variability implied by the standard deviation of the data-generating process, $\sigma_* = 50$ W/m^2 and yield $\bar{x} = 450$ W/m^2.

Modification of the prior distribution in light of the two annual data values using Bayes' Theorem yields the Gaussian posterior distribution in Equation 6.24, with posterior mean $\mu_h' = (550/212^2 + (2)(450)/50^2)/(1/212^2 + 2/50^2) = 453.4$ W/m^2, and posterior standard deviation $\sigma_h' = (1/212^2 + 2/50^2)^{-1/2} = 34.9$ W/m^2. The prior and posterior PDFs are compared in Figure 6.6. Having observed the wind power density for 2 years, uncertainty about its annual average value has decreased substantially. Even though the sample size of $n = 2$ is small, knowing that the generating-process standard deviation is 50 W/m^2, which is much smaller than the standard deviation of the prior distribution, has allowed these few data values to strongly constrain the location and spread of plausible values for μ in the posterior distribution. The probability, according to the posterior distribution, that the average annual wind power density is smaller than 400 W/m^2 is $\Pr\{z < (400 - 453.4)/34.9\} = \Pr\{z < -1.53\} = 0.063$.

The probability distribution for a future year's average wind power density, which would be if interest if the wind generation facility were to be built, is the Gaussian predictive distribution with parameters calculated using Equation 6.26, which are $\mu_+ = 453.4$ and $\sigma_+ = 61.0$ W/m^2. This distribution reflects uncertainty due both to the intrinsic interannual variability of the wind power density,

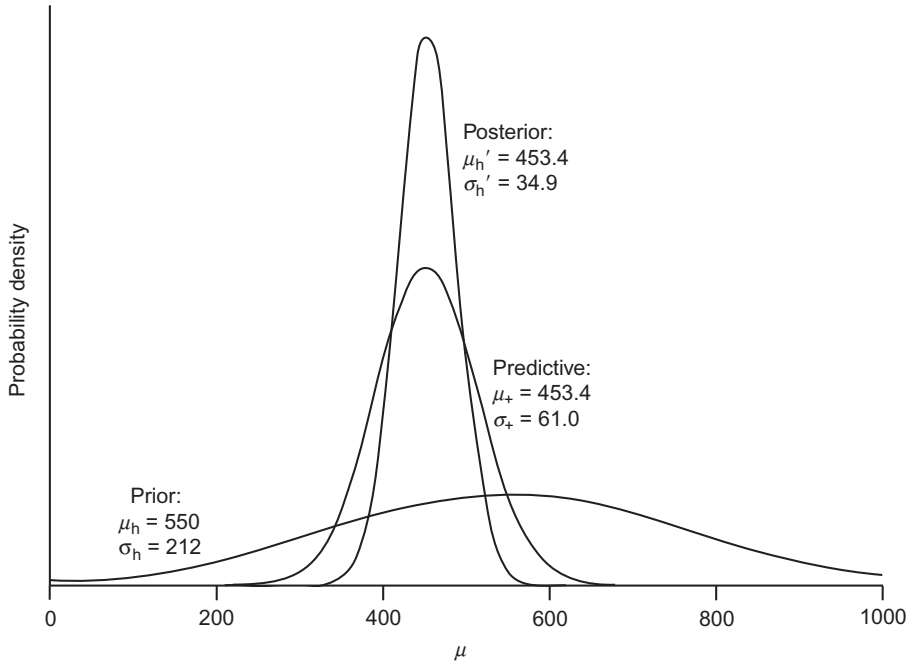

FIGURE 6.6 Prior, posterior, and predictive Gaussian distributions for the annually averaged wind power density, W/m^2.

characterized by σ_*^2, as well as uncertainty about the underlying climatological mean value μ that is expressed by the posterior distribution. ◇

6.4. DEALING WITH DIFFICULT INTEGRALS

6.4.1. Markov Chain Monte Carlo (MCMC) Methods

Not all data-generating processes can be characterized by likelihood functions having conjugate prior and posterior distributions. Nor is it always the case that the form of a conjugate prior distribution is capable of adequately representing an analyst's beliefs about the parameter or parameters of the data-generating process, so that a nonconjugate prior distribution must be used. In either of these cases, the normalizing integral in the denominator of Equation 6.1 may not exist in closed form, and its explicit numerical integration may be difficult. The same problems often occur for the integral of the posterior distribution, on the left-hand sides of Equations 6.1 and 6.2, evaluation of which are necessary for computation of inferential quantities such as credible intervals.

The usual approach to Bayesian inference in such settings is the use of *Markov chain Monte Carlo*, or MCMC, methods. Rather than attempting to compute explicit expressions for, or numerical approximations to, the relevant integrals, MCMC methods operate through statistical simulation, or generation of (pseudo-) random samples from the distributions of interest, using "Monte Carlo" methods of the kinds described in Section 4.7. MCMC algorithms yield sequences of simulated values from a target distribution that constitute what is called a Markov chain, which means that these sequences of random numbers are not independent but rather exhibit a particular form of serial dependence. Markov chains for sequences of discrete variables are discussed in Section 10.2.

Given two conditions that are usually met when using MCMC methods, namely, that the Markov chain is aperiodic (never repeats exactly) and irreducible (cannot reach a point where some of the allowable values can never again be simulated), a very large sample of these simulated values approaches the target distribution. If the target distribution from which the random values have been drawn is the posterior distribution for a Bayesian analysis, then attributes of this distribution (e.g., posterior moments, credible intervals, etc.) can be well approximated using sample counterparts from a large collection of simulated values.

Convergence of the empirical distribution of random values from a MCMC algorithm to the actual underlying distribution as $n \rightarrow \infty$ occurs even though these samples from the target distribution are not mutually independent. Therefore the serial correlation in the simulated values does not present a problem if we are interested only in computing selected quantiles or moments of the target distribution. However, if (approximately) independent samples from the target distribution are needed, or if computer storage must be minimized, the chain may be "thinned." *Thinning* simply means that most of the simulated values are discarded, and only every mth simulated value is retained. An appropriate value of m depends on the nature and strength of the serial correlation in the simulated values, and might be estimated using the variance inflation factor, or "time between effectively independent samples" in Equation 5.13. Because simulated MCMC sequences may exhibit quite strong serial correlation, appropriate values of m can be 100 or larger.

Another practical issue to be considered is ensuring the convergence of the simulated values to the target distribution. Depending on the value used to initialize a Markov chain, the early portion of a simulated sequence may not be representative of the target distribution. It is usual practice to discard this first portion of a simulated sequence, called the *burn-in* period. Sometimes the length of the burn-in period is chosen arbitrarily (e.g., discard the first 1000 values), although a better practice is to create at a scatterplot of the simulated values as a function of their position number in the sequence, and look for a place after which the point scatter appears to "level off" and fluctuate with unchanging variance around a fixed value. Similarly, it is good practice to ensure that the simulations are being generated from an irreducible Markov chain, by initializing multiple simulated sequences from different starting points, and checking that the resulting distributions are the same, following the burn-in period.

Two approaches to constructing MCMC sequences are in general use. These are described in the next two sections.

6.4.2. The Metropolis–Hastings Algorithm

The *Metropolis–Hastings algorithm* is a procedure for random number generation that is similar to the rejection method (Section 4.7.3). In both cases it is necessary to know only the mathematical form of the PDF of the target distribution, and not its CDF (so the PDF need not be analytically integrable). Also in common with the rejection method, candidates for the next simulated value are drawn from a different distribution that is easy to sample from, and each candidate value may be accepted or not, depending on an additional random draw from the uniform [0,1] distribution. The Metropolis–Hastings algorithm is especially attractive for Bayesian inference because only a function proportional to the target PDF (Equation 6.2) needs to be known, rather than the complete PDF of the posterior distribution (Equation 6.1). In particular, the integral in the denominator of Equation 6.1 need never be computed.

To simulate from a posterior distribution $f(\theta|x)$, it is first necessary to choose a candidate-generating distribution $g(\theta)$ that is easy to simulate from, and which has the same support as $f(\theta|x)$. That is, $g(\theta)$ and $f(\theta|x)$ must be defined over the same range of the random argument θ.

The Metropolis–Hastings algorithm begins by drawing a random initial value, θ_0, from $g(\theta)$ for which $f(\theta_0|x) > 0$. Then, for each iteration, i, of the algorithm a new candidate value, θ_C, is drawn from the candidate-generating distribution, and used to compute the ratio

$$R = \frac{f(\theta_C|x)/f(\theta_{i-1}|x)}{g(\theta_C)/g(\theta_{i-1})}, \tag{6.27}$$

where θ_{i-1} denotes the simulated value from the previous iteration. Notice that the target density $f(\theta|x)$ appears as a ratio in Equation 6.27, so that whatever the normalizing constant in the denominator of Equation 6.1 might be, it cancels in the numerator of Equation 6.27.

Whether or not the candidate value θ_C is accepted as the next value, θ_i, in the Markov chain depends on the ratio in Equation 6.27. It will be accepted if $R \geq 1$,

$$\text{For } R \geq 1,\ \theta_i = \theta_C, \tag{6.28a}$$

otherwise

$$\text{For } R < 1,\ \theta_i = \begin{cases} \theta_C & \text{if } u_i \leq R \\ \theta_{i-1} & \text{if } u_i > R \end{cases}. \tag{6.28b}$$

That is, if $R \geq 1$ then θ_C is automatically accepted as the next value in the chain. If $R < 1$, then θ_C is accepted if u_i, which is an independent draw from the uniform [0,1] distribution, is no greater than R. Importantly, and differently from the rejection method described in Section 4.7.3, the previous value θ_{i-1} is repeated if the candidate value is not accepted.

The algorithm based on the ratio in Equation 6.27 is called "independence" Metropolis–Hastings sampling, but the resulting sequence of simulated values $\theta_1, \theta_2, \theta_3, \ldots$ is nevertheless a Markov chain exhibiting serial correlation, and that serial correlation may be quite strong. The procedure generally works best if the candidate-generating distribution $g(x)$ has heavier tails than the target distribution, which suggests that the prior distribution $f(\theta)$ may often be a good choice for the candidate-generating distribution, particularly if a straightforward algorithm is available for simulating from it.

Example 6.5 Gaussian Inference Without a Conjugate Prior Distribution

Example 6.4 considered evaluation of a hypothetical site for its wind power potential, using a Gaussian data-generating function to represent interannual variations in wind power density, and a conjugate prior distribution with mean 500 W/m^2 and standard deviation 212 W/m^2. This formulation was convenient, but the Gaussian prior distribution might not adequately represent an evaluator's prior beliefs about the wind power potential of the site, particularly as this prior distribution specifies a small but nonzero (= 0.0048) probability of impossible negative wind power densities.

Alternatively, the analyst might prefer to use a functional form for the prior distribution with support only on the positive part of the real line, such as the Weibull distribution (Equation 4.69). If, as before, the median and 5th percentile of the analyst's subjective distribution are 550 and 200 W/m^2, respectively, Equation 4.70 can be used to find that the consistent Weibull distribution parameters are $\alpha = 2.57$ and $\beta = 634$ W/m^2.

The likelihood consistent with a Gaussian data-generating process is, as before, Equation 6.23b, and the Weibull prior distribution is proportional to

$$f(\mu) \propto \left(\frac{\mu}{\beta}\right)^{\alpha-1} \exp\left[-\left(\frac{\mu}{\beta}\right)^{\alpha}\right], \tag{6.29}$$

because the factor α/β in Equation 4.69 does not depend on μ. Accordingly, the posterior density is proportional to the product of Equations 6.23b and 6.29,

$$f(\mu|\,\overline{x}) \propto \exp\left[\frac{-n}{2\sigma_*^2}(\overline{x}-\mu)^2\right] \left(\frac{\mu}{\beta}\right)^{\alpha-1} \exp\left[-\left(\frac{\mu}{\beta}\right)^{\alpha}\right], \tag{6.30}$$

where as before $\sigma_* = 50\,\mathrm{W/m^2}$ is the known standard deviation of the Gaussian data-generating process, and the sample mean of $\overline{x} = 450\,\mathrm{W/m^2}$ was computed on the basis of $n = 2$ years of exploratory wind measurements.

The posterior PDF in Equation 6.30 is not a familiar form, and it is not clear that the normalizing constant (denominator in Equation 6.1) for it could be computed analytically. However, the Metropolis–Hastings algorithm allows simulation from this PDF, using a candidate-generating distribution $g(\mu)$ with the same support (positive real numbers) from which it is easy to simulate. A plausible choice for this candidate-generating distribution is the prior Weibull distribution $f(\mu)$, which clearly has the same support. Weibull variates can be generated easily using the inversion method (Section 4.7.4), as illustrated in Exercise 4.16.

Table 6.1 presents the results of the first 10 iterations of a realization of the Metropolis–Hastings algorithm. The algorithm has been initialized at $\mu_0 = 550$, which is the median of the prior distribution and which corresponds to nonzero density in the posterior distribution: $f(\mu_0|\overline{x}) = 0.00732$. The draw from the candidate-generating distribution on the first iteration is $\mu_C = 529.7$, yielding $R = 4.310$ in Equation 6.27, so that this candidate value is accepted as the simulated value for the first iteration, μ_1. This value becomes μ_{i-1} in the second iteration, in which the new candidate value $\mu_C = 533.6$ is generated. This value for the candidate in the second iteration yields $R = 0.773 < 1$, so it is necessary to generate the uniform [0,1] random number $u_2 = 0.3013$. Since $u_2 < R$ the candidate value is accepted

TABLE 6.1 Values for the Quantities in Equations 6.27 and 6.28, for the First 10 Iterations of a Realization of the Metropolis–Hastings Algorithm, Beginning With the Initial Value $\mu_0 = 550\,\mathrm{W/m^2}$

It., i	μ_{i-1}	μ_C	$f(\mu_C\|\overline{x})$	$f(\mu_{i-1}\|\overline{x})$	$g(\mu_C)$	$g(\mu_{i-1})$	R	u_i	μ_i
1	550.0	529.7	0.03170	0.00732	0.00163	0.00162	4.310	–	529.7
2	529.7	533.6	0.02449	0.03170	0.00163	0.00163	0.773	0.3013	533.6
3	533.6	752.0	0.00000	0.02449	0.00112	0.00163	0.000	0.7009	533.6
4	533.6	395.7	0.10889	0.02449	0.00144	0.00163	5.039	–	395.7
5	395.7	64.2	0.00000	0.10889	0.00011	0.00144	0.000	0.9164	395.7
6	395.7	655.5	0.00000	0.10889	0.00144	0.00144	0.000	0.4561	395.7
7	395.7	471.2	0.32877	0.10889	0.00160	0.00144	2.717	–	471.2
8	471.2	636.6	0.00000	0.32877	0.00149	0.00160	0.000	0.0878	471.2
9	471.2	590.0	0.00015	0.32877	0.00158	0.00160	0.000	0.4986	471.2
10	471.2	462.3	0.36785	0.32877	0.00158	0.00160	1.128	–	462.3

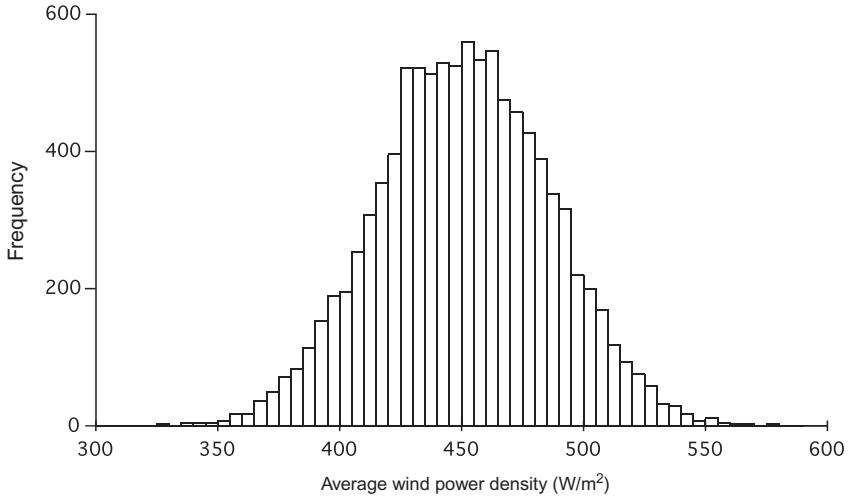

FIGURE 6.7 Histogram of 10,000 random draws from the posterior distribution in Equation 6.30, generated by the Metropolis–Hastings algorithm. The mean and standard deviation of this distribution are 451.0 and 35.4 W/m^2, respectively.

as $\mu_2 = 533.6$. In the third iteration, the candidate value of 752.0 is an extreme tail value in the posterior distribution, which yields $R = 0.000$ (to three decimal places). Since $u_3 = 0.7009 > R$, the candidate value for the third iteration is rejected, and the generated value is the same as that from the second iteration, $\mu_3 = \mu_2 = 533.6$.

The process begun in Table 6.1 can be continued indefinitely, and for this simple example the necessary computations are very fast. Figure 6.7 shows a histogram of 10,000 of the resulting values generated from the posterior distribution, which are the results of every $m = 100$th of 1,000,000 iterations. Since the Weibull prior distribution used to arrive at this posterior distribution is very similar to the Gaussian prior distribution shown in Figure 6.6, it is not surprising that the histogram in Figure 6.7 is similar to the posterior distribution in Figure 6.6. The mean and standard deviation of the histogram in Figure 6.7 are 451.0 and 35.4 W/m^2, which are similar to the mean and standard deviation of 453.4 and 34.9 W/m^2, respectively, of the posterior distribution in Figure 6.6. $\Pr\{\mu < 400\,\text{W/m}^2\} = 0.076$ according to Figure 6.7, as compared to $\Pr\{\mu < 400\,\text{W/m}^2\} = 0.063$ for the posterior distribution in Figure 6.6.

The result in Figure 6.7 was produced with essentially no burn-in, other than having discarded results from the first $m - 1 = 99$ iterations. However, a scatterplot of the 10,000 values in Figure 6.7 as a function of their iteration number showed no apparent trends, either in location or dispersion. ◇

6.4.3. The Gibbs Sampler

The Metropolis–Hastings algorithm is usually the method of choice for MCMC Bayesian inference in one-parameter problems, when a prior distribution conjugate to the form of the data-generating process is either not available or not suitable. It can also be implemented in higher-dimensional problems (i.e., those involving simultaneous inference about multiple parameters) when an appropriate higher-dimensional candidate-generating distribution is available. However, when simultaneous inferences regarding two or more parameters are to be computed, an alternative MCMC approach called the *Gibbs*

sampler is more typically used. Casella and George (1992) present a gentle introduction to this algorithm.

The Gibbs sampler produces samples from a K-dimensional posterior distribution, where K is the number of parameters being considered, by simulating from the K univariate conditional distributions for each of the parameters, given fixed values for the remaining $K-1$ parameters. That is, a given K-dimensional joint posterior distribution $f(\theta_1, \theta_2, \theta_3, \ldots, \theta_K|x)$ can be characterized using the K univariate conditional distributions $f(\theta_1|\theta_2, \theta_3, \ldots, \theta_K, x), f(\theta_2|\theta_1, \theta_3, \ldots, \theta_K, x), \ldots, f(\theta_K|\theta_2, \theta_3, \ldots, \theta_{K-1}, x)$. Simulating from these individually will generally be easier and faster than simulating from the full joint posterior distribution. Denoting the simulated value for the kth parameter on the ith iteration as $\theta_{i,k}$, the ith iteration of the Gibbs sampler consists of the K steps:

1. Generate $\theta_{i,1}$ from $f(\theta_1|\theta_{i-1,2}, \theta_{i-1,\,3}, \ldots, \theta_{i-1,K}, x)$
2. Generate $\theta_{i,2}$ from $f(\theta_2|\theta_{i,1}, \theta_{i-1,\,3}, \ldots, \theta_{i-1,K}, x)$
⋮
k. Generate $\theta_{i,k}$ from $f(\theta_k|\theta_{i,1}, \theta_{i,2}, \theta_{i,k-1}, \ldots\ \theta_{i-1,\,k+1}, \ldots, \theta_{i-1,K}, x)$
⋮
K. Generate $\theta_{i,K}$ from $f(\theta_K|\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,K-1}, x)$

The ith realization for θ_1 is simulated, conditional on values for the other $K-1$ parameters generated on the previous, $(i-1)^{st}$, iteration. The ith realization for θ_2 is simulated conditionally on the value $\theta_{i,1}$ just generated, and values for the remaining $K-2$ parameters from the previous iteration. In general, for each step within each iteration, values for the conditioning variables are the ones that have most recently become available. The procedure begins with initial ("0th iteration") values $\theta_{0,1}, \theta_{0,2}, \theta_{0,3}, \ldots, \theta_{0,K}$, drawn perhaps from the prior distribution.

Occasionally, analysis of the joint posterior distribution $f(\theta_1, \theta_2, \theta_3, \ldots, \theta_K|x)$ may yield explicit expressions for the K conditional distributions to be simulated from. More typically, Gibbs sampling is carried out numerically using freely available software such as BUGS (Bayesian inference Using Gibbs Sampling), or JAGS (Just Another Gibbs Sampler), which can be found through web searches on these acronyms. Regardless of whether the K conditional distributions are derived analytically or evaluated with software, the results are serially correlated Markov chains for simulated values of the parameters θ_k. The same burn-in and possible thinning considerations discussed in the previous sections are applicable to Gibbs samplers as well.

Gibbs sampling is especially well suited to Bayesian inference for *hierarchical models*, where the hyperparameters of a prior distribution are themselves endowed with their own prior distributions, called *hyperpriors*. Such models arise naturally when the parameter(s) of the data-generating process depend on yet other parameters that are not themselves explicit arguments of the likelihood.

Example 6.6 Hierarchical Bayesian Model for Hurricane Occurrences

Elsner and Jagger (2004) have investigated the relationship between annual numbers of U.S. landfalling hurricanes and two well-known features of the climate system, using a hierarchical Bayesian model. The first of these features is the El Niño-Southern Oscillation (ENSO) phenomenon, which they represented using the "cold tongue index", or average sea-surface temperature anomaly in the equatorial Pacific region bounded by 6°N–6°S and 180°–90°W. The second of these features is the *North Atlantic Oscillation* (NAO), which is represented by an index reflecting the strength and orientation of the pair of mutual teleconnectivity features over the Atlantic Ocean in Figure 3.33.

The data-generating process responsible for the number of hurricanes, x_i, in year i is assumed to be Poisson, with mean μ_i that may be different from year to year, depending on the state of the climate system as represented in terms of indices of ENSO and NAO,

$$\ln(\mu_i) = \beta_0 + \beta_1 CTI_i + \beta_2 NAO_i + \beta_3 CTI_i NAO_i. \tag{6.31}$$

This hierarchical model is a Bayesian Poisson regression model, similar to the Poisson regression solved using maximum likelihood in Section 7.6.3. The logarithmic transformation on the left-hand side of Equation 6.31 ensures that the modeled μ_i will be strictly positive, as required. The resulting likelihood for the data-generating function, including the implicit expression for the μ_i in Equation 6.31, is (compare Equations 6.14 and 6.17)

$$f(x|\,\beta_0,\beta_1,\beta_2,\beta_3) \propto \prod_{i=1}^{n} \{\,[\exp(\beta_0 + \beta_1 CTI_i + \beta_2 NAO_i + \beta_3 CTI_i NAO_i)]^{x_i} \times \exp[-\exp(\beta_0 + \beta_1 CTI_i + \beta_2 NAO_i + \beta_3 CTI_i NAO_i)]\,\}. \tag{6.32}$$

Inferences in this hierarchical model focus on the posterior distributions for the β's, and begin with specification of a (hyper-) prior distribution for them. The multivariate normal distribution (Equation 12.1) is a straightforward and usual choice in models like this, which characterizes initial uncertainty about each β individually as a distinct Gaussian distribution. Elsner and Jagger (2004) considered both a vague prior, and an informative prior based on 19th-century hurricane counts (as in Example 6.3).

Equation 6.31 is a complicated function, and when it is multiplied by the prior distribution (Equation 12.1) it yields an even more complicated posterior distribution for the four β's. However, simulations from it can be made using Gibbs sampling, and these were generated using BUGS. Using data for U.S. landfalling hurricane numbers, CTI, and NAO for the years 1900–2000, and vague priors for the four β's, Elsner and Jagger (2004) simulated the marginal posterior distributions for them in Figure 6.8. These are actually kernel density estimates (Section 3.3.6) computed with Gaussian kernels and smoothing parameter 0.17.

The posterior means and standard deviations in Figure 6.8 are (a) −0.380 and 0.125, (b) −0.191 and 0.078, and (c) 0.200 and 0.102. Panels (a) and (b) in Figure 6.8 suggest strongly that average annual U.S. landfalling hurricane numbers are meaningfully related to both CTI (more landfalling hurricanes on average for negative CTI, or La Niña conditions) and NAO (more U.S. landfalling hurricanes on average for negative NAO, or relatively lower pressures in the subtropical Atlantic), since in both cases values

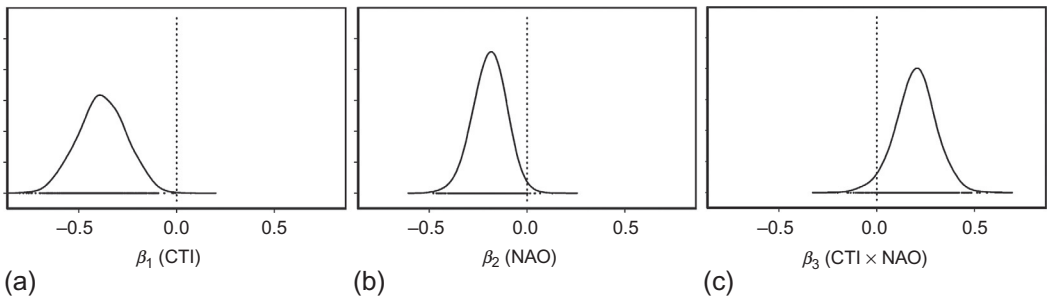

FIGURE 6.8 Marginal posterior distributions for the parameters (a) β_1, (b) β_2, and (c) β_3 in Equation 6.31. *From Elsner and Jagger (2004). © American Meteorological Society. Used with permission.*

near zero are unlikely and there is nearly zero probability that either coefficient is positive. The corresponding inference for β_3 in Figure 6.8c is not as strong, but assuming an approximately Gaussian shape for this posterior distribution implies the estimate $\Pr\{\beta_3 \le 0\} \approx \Pr\{z \le -2.00/0.102\} = \Pr\{z \le -1.96\} = 0.025$. ◇

6.5. EXERCISES

6.1 Suppose a different analyst considering the data in Example 6.2 concludes that a reasonable prior distribution for the binomial p in this situation is Gaussian, with mean 2/3 and standard deviation 1/10.
   a. Find the parameters of a beta distribution that approximates this Gaussian prior distribution.
   b. Using the results of part (a), find the posterior distribution for p.
   c. Find the resulting predictive distribution for the number of "successes" in the next $N^+ = 5$ independent observations. (Use a computer to calculate the logs of the gamma function.)

6.2 Suppose you have concluded that your prior distribution for a parameter of interest is well represented by a Gumbel distribution. Evaluate the parameters of this distribution if
   a. The interquartile range of your prior distribution is (100, 400).
   b. The mean and standard deviation of your prior distribution are 270 and 200, respectively.
   c. What do these two distributions imply about your beliefs about the magnitude of the 100-year event?

6.3 Assume the annual numbers of tornados occurring in a particular county is well described by the Poisson distribution. After observing two tornados in this county during 10 years, a Bayesian analysis yields a posterior distribution for the Poisson rate that is a gamma distribution, with $\alpha' = 3.5$ and $\beta' = 0.05$.
   a. What was the prior distribution?
   b. What is the probability of the county experiencing at least one tornado next year?

6.4 Recalculate Example 6.4 if the analyst has less uncertainty about the eventual suitability of the site for wind power generation, so that an appropriate prior distribution is Gaussian with mean $\mu_h = 550\,\mathrm{W/m^2}$ and standard deviation $\sigma_h = 100\,\mathrm{W/m^2}$.
   a. Find the posterior distribution.
   b. Find the predictive distribution.

6.5 Consider how the analysis in Example 6.4 would change if a third year of wind measurements had been obtained, for which the average annual wind power density was $375\,\mathrm{W/m^2}$.
   a. Find the updated posterior distribution.
   b. Find the updated predictive distribution.

6.6 What value would be generated for μ_{11} in Table 6.1 after the 11th iteration if.
   a. $\mu_C = 350$ and $u_{11} = 0.135$?
   b. $\mu_C = 400$ and $u_{11} = 0.135$?
   c. $\mu_C = 450$ and $u_{11} = 0.135$?