6

An Introduction to Bayesian Estimation

6.1 CHAPTER OVERVIEW

Chapter 7 introduces a second “modern” missing data approach, multiple imputation. Multiple imputation generates several copies of the data set and fills in (i.e., imputes) each copy with different estimates of the missing values. The imputation process is conceptually straightforward because it closely resembles the stochastic regression procedure from Chapter 2 (i.e., impute missing values with predicted scores and add a random residual to each imputed value). However, the mathematical machinery behind multiple imputation is heavily entrenched in Bayesian methodology. At one level, it is possible to effectively implement multiple imputation in a research study without fully understanding its Bayesian underpinnings. For example, multiple imputation software packages employ default settings that make Bayesian aspects of the analysis transparent to the user, and many multiple imputation primers make little to no reference to Bayesian methodology (Allison, 2002; Enders, 2006; Schafer & Graham, 2001; Schafer & Olsen, 1998; Sinharay, Stern, & Russell, 2001). However, understanding multiple imputation at a deeper level requires a background in Bayesian statistics, and accessing the seminal missing data work (Little & Rubin, 2002; Rubin, 1987; Schafer, 1997) can be difficult without this knowledge.

This chapter takes a hiatus from missing data issues to focus on Bayesian estimation. The goal of the chapter is to provide a user-friendly introduction to Bayesian statistics, while still providing a level of detail that will serve as a springboard for accessing the technically oriented missing data literature. The chapter is far from comprehensive, and I focus on aspects of Bayesian estimation that are particularly relevant to a multiple imputation analysis. A number of comprehensive resources are available in the literature (e.g., Bolstad, 2007; Gelman, Carlin, Stern, & Rubin, 1995), as are additional primer articles (e.g., Lee & Wagenmakers, 2005; Pruzek, 1997; Rupp, Dey, & Zumbo, 2004; Stephenson & Stern, 1998).

6.2 WHAT MAKES BAYESIAN STATISTICS DIFFERENT?

The definition of a parameter is a key distinction between Bayesian estimation and the so-called **frequentist paradigm** that is the predominant approach to estimation and significance testing in many disciplines (e.g., psychology, education, business). The frequentist approach defines a parameter as a fixed characteristic of the population. The goal of a frequentist analysis is to estimate *the* true value of the parameter and establish a confidence interval around that estimate. The standard error is integral to this process and estimates the variability of the estimate across repeated samples. Defining a parameter as a fixed quantity leads to some important subtleties. For example, consider the interpretation of a 95% confidence interval. It is incorrect to say that there is a 95% probability that the parameter falls between values of A and B because the confidence interval from any single sample contains the parameter or it does not. Instead, the confidence interval describes the expected performance of the interval across repeated samples. For example, if you drew 100 samples from a population and constructed a 95% confidence interval around the parameter estimate from each sample, you would expect 95 of the intervals to include the population parameter. In a similar vein, the probability value from a frequentist significance test describes the proportion of repeated samples that would yield a test statistic equal to or greater than that of the data. In both situations, the probability statement applies to the *data*, not to the parameter.

In contrast, the **Bayesian paradigm** views a parameter as a random variable that has a distribution. One of the goals of a Bayesian analysis is to describe the shape of this distribution. For example, the mean and the standard deviation describe the distribution's center and spread, respectively. The mean quantifies the parameter's most likely value (assuming that the distribution is symmetric) and is similar to a frequentist point estimate. The standard deviation (or alternatively, the variance) is analogous to a frequentist standard error, but it describes the degree of uncertainty about the parameter after observing the data. The Bayesian notion of uncertainty does not involve repeated sampling. Viewing the parameter as a random variable contrasts the frequentist approach in other ways. For example, a Bayesian **credible interval** (the analog to a frequentist confidence interval) allows you to say that there is a 95% probability that the parameter falls between values of A and B. This interpretation is very different from that of the frequentist approach because it attaches the probability statement to the *parameter*, not to the data.

6.3 A CONCEPTUAL OVERVIEW OF BAYESIAN ESTIMATION

A Bayesian analysis consists of three major steps: (1) specify a prior distribution for the parameter of interest, (2) use a likelihood function to summarize the data's evidence about different parameter values, and (3) combine information from the prior distribution and the likelihood to generate a posterior distribution that describes the relative probability of different parameter values. Describing the shape of the posterior distribution is a key goal of a Bayesian analysis, and familiar statistics such as the mean and the variance summarize the location (i.e., the center of) and the spread of the posterior, respectively. This section gives a conceptual description of these three steps. Because the goal is to introduce the underlying

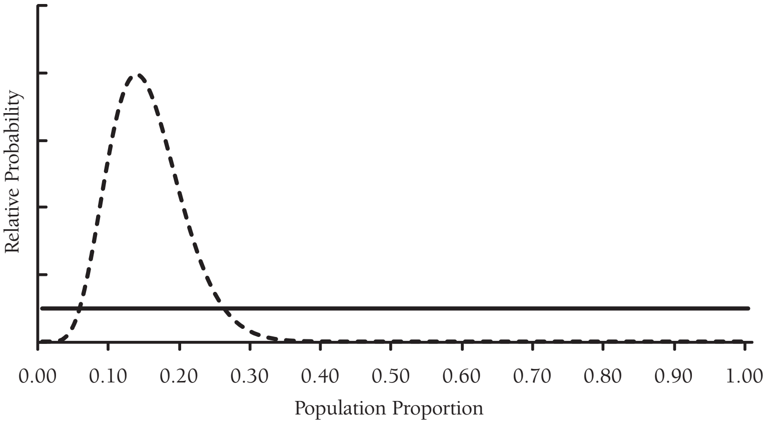

FIGURE 6.1. The prior distributions from the depression example. Researcher A's prior is the dashed curve that assigns higher probabilities to population proportions between 0.10 and 0.15. Researcher B's prior distribution is the solid line that assigns an equal weight to every parameter value.

logic behind Bayesian estimation, I am purposefully vague about many of the mathematical details. For now, I use a straightforward univariate analysis example where the goal is to estimate the proportion of clinically depressed individuals in a population. Subsequent sections, however, give a more thorough description of the mathematics and illustrate the application of Bayesian estimation to a mean vector and a covariance matrix (the key parameters in a multiple imputation analysis). As you will see, multivariate analyses use the same three-step procedure described in this section.

The Prior Distribution

The first step in a Bayesian analysis is to specify a prior distribution for the parameter of interest. The **prior distribution** describes your subjective beliefs about the relative probability of different parameter values before collecting any data. To illustrate, suppose that two researchers want to use Bayesian methodology to estimate the proportion of clinically depressed individuals in a population, π. The prior distribution specifies the relative probability of every possible population proportion. After conducting a literature review, Researcher A believes that depression rates between 0.10 and 0.15 are very likely, and she feels that the relative probability rapidly decreases as the proportion approaches zero or one. The dashed curve in Figure 6.1 depicts this researcher's prior beliefs. Notice that the highest point of the distribution is located at $\pi = 0.13$, and the relative probability (i.e., the height of the curve) quickly decreases as π approaches zero or one. In contrast, Researcher B is uncomfortable speculating about different parameter values, so he assigns an equal weight to every proportion between zero and one. The flat line in Figure 6.1 depicts this researcher's prior beliefs. The Bayesian literature often refers to Researcher B's prior distribution as a **noninformative prior** because it represents a lack of knowledge about the population parameter.

The Likelihood Function

The second step of a Bayesian analysis is to collect data and use a **likelihood function** to summarize the data's evidence about different parameter values. This step applies the maximum

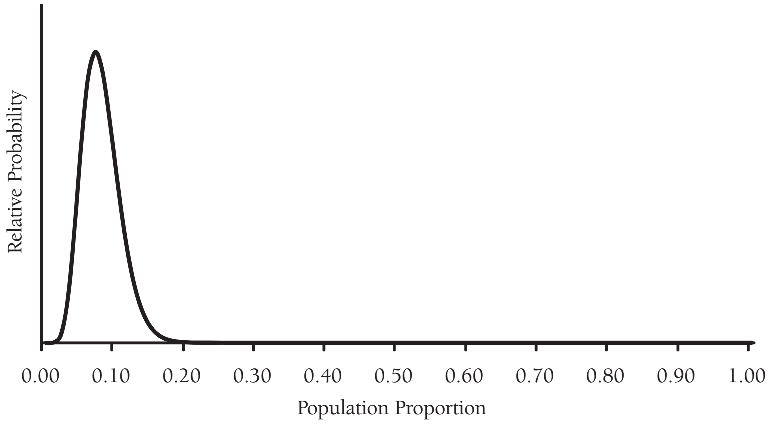

FIGURE 6.2. The binomial likelihood function from the depression example. The height of the likelihood function gives the relative probability that the population on the horizontal axis would produce a sample where 7 out of 100 individuals are diagnosed with depression. The maximum of the function (i.e., the maximum likelihood estimate) corresponds with π = 0.07, which is the sample proportion.

likelihood principles that I outlined in Chapter 3, but uses the likelihood rather than the log-likelihood. Recall from Chapter 3 that substituting the data and a parameter value into a probability density function (e.g., the equation that defines the normal curve) returns the likelihood (i.e., relative probability) of the data, given that particular parameter value. Repeating this process for different parameter values yields a likelihood function that describes the relative probability of the data across a range of parameter values.

For example, suppose that the two researchers drew a sample and found that 7 out of the 100 individuals whom they assessed met their criteria for clinical depression. The binomial density function is the appropriate likelihood for a binary outcome variable. The binomial density function is quite different from that of the normal curve in Chapter 3, but it works in the same way. Specifically, you substitute the data (e.g., 7 out of 100 diagnosed cases) and a population proportion (e.g., $\pi = 0.15$) into the density function, and the equation returns the likelihood of observing the sample data from a population with that particular prevalence rate. Repeating the computations using different population proportions yields a likelihood function that shows how the probability of the data varies as a function of π. For example, Figure 6.2 shows the binomial likelihood function for the depression data. Notice that the maximum likelihood estimate (i.e., the highest point on the function) is the sample proportion, $\hat{\pi} = .07$.

The Posterior Distribution

The final step of a Bayesian analysis is to define the posterior distribution of the parameter. The **posterior distribution** is a composite distribution that combines information from the prior and the likelihood to generate an updated set of relative probabilities. I describe the posterior in more detail later in the chapter, but the basic idea is to weight each point on the likelihood function by the magnitude of your prior beliefs. For example, if you attached a high prior probability to a particular parameter value, the posterior would increase the height of the likelihood function at that point on the horizontal axis. Conversely, if you assigned a low prior probability to a particular parameter value, the posterior would decrease the height of the likelihood function at that point.

To illustrate, reconsider the depression scenario. Prior to collecting data, Researcher A assigned a high probability to depression rates between 0.10 and 0.15. The data supported somewhat lower values and indicated that $\pi = 0.07$ is the most likely population proportion. Figure 6.3 shows Researcher A's posterior distribution as a dashed line. The effect is subtle, but her posterior distribution is a blend of her prior and the likelihood function. For reasons that I explain later, the solid line in Figure 6.3 (Researcher B's posterior distribution) is identical to the likelihood function. Comparing the relative height of the two curves at $\pi = 0.05$, you can see that Researcher A's posterior distribution is less elevated than the likelihood function. Researcher A assigned a very low prior probability to $\pi = 0.05$, which effectively downweights the likelihood function at that point. Next, compare the relative height of the two distributions at $\pi = .15$. Researcher A assigned a high prior probability to this parameter value, so her posterior distribution is slightly elevated relative to the likelihood function (i.e., the prior probability boosts this point on the likelihood function). In contrast, Researcher B specified a prior distribution where every parameter value has the same probability. Consequently, his posterior distribution weights every point on the likelihood function by the same amount and is identical to the likelihood function in Figure 6.2.

Summarizing the shape of the posterior distribution is an important part of a Bayesian analysis. Without delving into any equations, Researcher A's posterior distribution has a mean of 0.095, a mode of 0.090, and a standard deviation of 0.024. In contrast, Researcher B's posterior has a mean of 0.078, a mode of 0.070, and a standard deviation of 0.026. The fact that Researcher A's distribution has somewhat higher measures of central tendency follows from the fact that she assigned high prior probabilities to proportions between 0.10 and

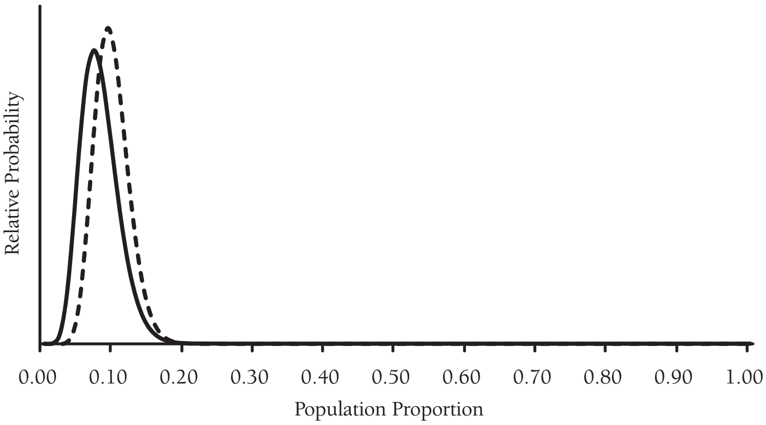

FIGURE 6.3. The posterior distributions from the depression example. Researcher A's posterior distribution is the dashed curve. She specified a prior distribution that assigns higher weights to population proportions between 0.10 and 0.15. Consequently, her posterior distribution has shifted slightly to the right of the likelihood function in Figure 6.2. Researcher B's posterior distribution is the solid curve. He specified a prior distribution where every parameter value has the same probability, so his posterior distribution is identical to the likelihood.

0.15. Nevertheless, the relative similarity of the two sets of summary statistics is noteworthy, particularly given that the researchers adopted radically different prior distributions.

For comparison purposes, a frequentist analysis of the depression data yields a point estimate and standard error of $\hat{\pi} = 0.07$ and $SE = 0.026$, respectively. Notice that these estimates are identical to Researcher B's posterior mode and posterior standard deviation. Although the Bayesian and frequentist analyses give the same numeric results, they have very different interpretations. For example, $\hat{\pi}$ estimates *the* true population proportion, and the standard error quantifies the variability of the point estimate across repeated samples. In contrast, the posterior mode is the most likely value from a distribution of proportions, and the posterior standard deviation quantifies the spread of the parameter distribution.

More on the Prior Distribution

A Bayesian analysis uses the prior distribution to incorporate subjective beliefs as a data source. This may be troublesome to researchers who are accustomed to the frequentist paradigm, but the idea of using prior information actually makes good intuitive sense. For example, suppose that a researcher had access to a meta-analysis prior to designing a study. Meta-analyses estimate the average effect size in a body of research and often summarize the variability of the effect across different design characteristics (e.g., Ioannidis et al., 2001; Lipsey & Wilson, 1993; Rubin, 1992). The Bayesian approach provides a mechanism for incorporating prior knowledge into an analysis (e.g., by using the meta-analysis to formulate a prior distribution), whereas frequentist estimation essentially ignores the fact that previous studies even exist. In the frequentist paradigm, the benefit of having a meta-analysis is limited to estimating power, determining sample size, and formulating a directional hypothesis.

If the notion of using prior information as a data source still feels uncomfortable, there is one final consideration. The depression example did not illustrate this point, but it ends up that you can specify the amount of influence that the prior exerts on the analysis results. Specifying a prior distribution generally requires three pieces of information: the location of the distribution (e.g., its mean), the spread of the distribution (e.g., its standard deviation), and the number of "hypothetical data points" associated with the prior. Collectively, Bayesian texts sometimes refer to these characteristics as the distribution's **hyperparameters**. Importantly, you can use the sample size metric to quantify the prior distribution's influence. For example, if you have relatively little confidence in the prior distribution, you can assign a small number of imaginary data points to the prior. In contrast, you can assign a large number of data points to the distribution if you are very confident in your prior beliefs.

Returning to the depression example, note that the researchers' prior distributions have very different hyperparameters. The two distributions in Figure 6.1 are shaped quite differently, which implies that they differ with respect to their location and spread. However, the fact that the prior distributions imply different sample sizes is not so obvious. Without going into the mathematical details, Researcher A's prior distribution (i.e., the dashed curve) assigns a weight that is equivalent to approximately 45 imaginary data points. Because the prior is contributing roughly half as much information as the data, the resulting posterior distribution is a blend of the prior and the likelihood function. In contrast, Researcher B's non-informative prior distribution contributes nothing to the estimation process, so his posterior distribution has the same shape as the likelihood function, and his posterior mode is identical to the sample proportion. In general, adopting a noninformative prior yields a posterior distribution that is defined solely by the data. This is an important point that will be revisited in this chapter and the next.

6.4 BAYES' THEOREM

This section fills in some of the mathematical details omitted from the previous depression example. As you will see, Bayes' theorem is the mathematical machinery behind a Bayesian analysis and plays a key role in defining the shape of the posterior distribution. In fact, the three steps in a Bayesian analysis (i.e., specify a prior, estimate the likelihood, define the posterior) are terms in the theorem equation.

Bayes' theorem describes the relationship between two conditional probabilities. For two random events, A and B, the theorem is

$$p(B|A) = \frac{p(B)p(A|B)}{p(A)} \tag{6.1}$$

where $p(B|A)$ is the conditional probability of observing event B, given that event A has already occurred, $p(A|B)$ is the conditional probability of A given B, $p(B)$ is the probability of B alone, and $p(A)$ is the marginal probability of A.

The generic notation in Equation 6.1 offers little insight into the application of Bayes' theorem to statistics, but the linkage becomes slightly clearer if you replace A with the sample data and B with a parameter, as follows.

$$p(\theta|Y) = \frac{p(\theta)p(Y|\theta)}{p(Y)} \tag{6.2}$$

The terms in Equation 6.2 now align with the concepts that I introduced in the previous section. Specifically, θ is the parameter of interest (e.g., the proportion of clinically depressed individuals), Y is the sample data, $p(\theta)$ is the parameter's prior distribution, $p(Y|\theta)$ is the likelihood (i.e., the conditional probability of the data, given some assumed value of θ), $p(Y)$ is the marginal distribution of the data, and $p(\theta|Y)$ is the posterior distribution (i.e., the conditional probability of the parameter, given the data).

In words, Bayes' theorem is

$$\text{Posterior} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Scaling factor}} \tag{6.3}$$

I previously described the posterior distribution as a weighted likelihood function, where the basic idea is to adjust each point on the likelihood function by the magnitude of the corresponding prior probability. This is accomplished in the numerator of Bayes' theorem by multiplying the likelihood function by the corresponding prior probabilities. As I explain later, the denominator of the theorem is simply a scaling constant that makes the area under the posterior distribution sum (i.e., integrate) to one. Dividing by a constant does not change the basic shape of the posterior distribution, so ignoring the denominator yields the following simplified expression.

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood} \tag{6.4}$$

Equation 6.4 says that *the posterior distribution is proportional to the prior distribution times the likelihood*. This is the fundamental idea behind Bayesian estimation and is a point that will resurface throughout the rest of the chapter.

6.5 AN ANALYSIS EXAMPLE

Having filled in some of the mathematical details, I return to the depression example and illustrate how Bayes' theorem applies to a statistical analysis. Again, the basic procedure that I describe in this section generalizes to multivariate estimation problems and to a multiple imputation analysis.

The Prior Distribution

The first step of a Bayesian analysis is to specify a prior distribution. The prior distributions in Figure 6.1 belong to the **beta distribution** family (by family of distributions, I mean a collection of distributions that share the same basic shape or function, much like the *t*-distribution family). Like the normal curve, a probability density function defines the shape of the beta distribution. The beta density function is

$$p(\pi) \propto \pi^{a-1}(1-\pi)^{b-1} \tag{6.5}$$

where $p(\pi)$ is the height of the curve at a particular value of π, and a and b are constants that define the shape of the distribution (e.g., larger values of a and b produce a distribution with greater spread, and the distribution becomes asymmetric when $a \neq b$). Density functions typically contain a collection of scaling terms that make the area under the distribution sum to one. Excluding these terms has no bearing on the distribution's shape, so I omit the scaling factor from Equation 6.5 and use the "proportional to" symbol (i.e., $\propto$) to indicate that the two sides of the equation differ by a multiplicative constant. To simplify things, I use this convention throughout the chapter.

Returning to the depression example, note that Table 6.1 gives the height of the prior distributions at integer values of π between 0.05 and 0.20. To begin, consider the height of Researcher A's prior distribution at $\pi = 0.05$ and $\pi = 0.10$. Her prior is a beta distribution with $a = 7$ and $b = 40$, so substituting $\pi = 0.05$ and $\pi = 0.10$ into the beta density function yields values of 0.792 and 6.153, respectively (note that I used the previously omitted scaling constant in these calculations). Visually, 0.792 and 6.153 represent the height of the prior distribution at parameter values of $\pi = 0.05$ and $\pi = 0.10$, respectively. Similar to the likelihood values from Chapter 3, you can think of these quantities as relative probabilities. The relative magnitude of the prior probabilities reflects Researcher A's belief that $\pi = 0.10$ was a more

TABLE 6.1. Prior Distributions, Likelihood, and Posterior Distributions from the Depression Example

	Researcher A				Researcher B			
π	Prior	Likelihood	Prior × Likelihood	Scaled posterior	Prior	Likelihood	Prior × Likelihood	Scaled posterior
0.05	0.7919	0.1060	0.0840	0.0036	1.0000	0.1060	0.1060	0.1169
0.06	1.5651	0.1420	0.2222	0.0190	1.0000	0.1420	0.1420	0.1565
0.07	2.6006	0.1545	0.4018	0.0570	1.0000	0.1545	0.1545	0.1703
0.08	3.8012	0.1440	0.5472	0.1134	1.0000	0.1440	0.1440	0.1587
0.09	5.0318	0.1188	0.5979	0.1640	1.0000	0.1188	0.1188	0.1310
0.10	6.1533	0.0889	0.5470	0.1835	1.0000	0.0889	0.0889	0.0980
0.11	7.0505	0.0613	0.4321	0.1661	1.0000	0.0613	0.0613	0.0675
0.12	7.6483	0.0394	0.3013	0.1257	1.0000	0.0394	0.0394	0.0434
0.13	7.9170	0.0238	0.1887	0.0815	1.0000	0.0238	0.0238	0.0263
0.14	7.8679	0.0137	0.1075	0.0461	1.0000	0.0137	0.0137	0.0151
0.15	7.5429	0.0075	0.0563	0.0232	1.0000	0.0075	0.0075	0.0082
0.16	7.0027	0.0039	0.0273	0.0104	1.0000	0.0039	0.0039	0.0043
0.17	6.3153	0.0020	0.0124	0.0043	1.0000	0.0020	0.0020	0.0022
0.18	5.5466	0.0009	0.0052	0.0016	1.0000	0.0009	0.0009	0.0010
0.19	4.7543	0.0004	0.0021	0.0005	1.0000	0.0004	0.0004	0.0005
0.20	3.9842	0.0002	0.0008	0.0002	1.0000	0.0002	0.0002	0.0002
		Sums =	3.5338	1.0000			0.9073	1.0000

plausible parameter value than $\pi = 0.05$. Next, consider Researcher B's prior, which is a beta distribution with $a = 1$ and $b = 1$. In this situation, the beta density function in Equation 6.5 always returns a value of 1.00, so Researcher B is assigning the same weight to every possible value of π.

The Likelihood Function

The second step of a Bayesian analysis is to collect data and use a likelihood function to summarize the data's evidence about different parameter values. This step applies the maximum likelihood principles outlined in Chapter 3. Specifically, substituting the sample data and a parameter value (i.e., π) into a density function yields the likelihood (i.e., relative probability) of the data, given that parameter value. Repeating this process for different parameter values yields a likelihood function that describes the relative probability of the data across a range of parameter values. The **binomial density function** is the appropriate likelihood for a binary outcome variable (i.e., each individual is classified as depressed or not depressed). The binomial density function is

$$p(y\,|\,\pi) \propto \pi^{y}(1-\pi)^{N-y} \tag{6.6}$$

where $p(y\,|\,\pi)$ is the height of the curve at a particular value of π, y is the number of "successes" (e.g., the number of depressed individuals), and N is the total number of "trials" (e.g., the sample size). Again, I omit the scaling constant from the equation to simplify things.

Returning to the depression example, the researchers assessed a sample of 100 individuals and found that seven people met their criteria for clinical depression. Substituting $y = 7$ and $N = 100$ into Equation 6.6 yields the binomial likelihood function in Figure 6.2. The height of the likelihood function gives the relative probability of observing 7 depressed cases in a sample of 100 individuals, given the population parameter value on the horizontal axis (i.e., the conditional probability of the data, given some assumed value of π). Table 6.1 gives the numeric value of the likelihood for parameter values between $\pi = 0.05$ and 0.20 (again, I used the previously omitted scaling constant for these calculations in order to avoid excessive decimals). Consider the likelihood associated with $\pi = 0.05$ and $\pi = 0.10$, the values of which are 0.106 and 0.089, respectively. Consistent with the interpretation of the likelihood in Chapter 3, 0.106 and 0.089 are the relative probabilities of observing the data (i.e., 7 out of 100 diagnosed individuals) from a population with $\pi = 0.05$ and $\pi = 0.10$, respectively. Visually, these numeric values correspond with the height of the curve at $\pi = 0.05$ and $\pi = 0.10$. Because $\pi = 0.05$ returns a higher relative probability than $\pi = 0.10$, the data provide slightly more evidence in favor of $\pi = 0.05$.

Before proceeding, you may have noticed that Equations 6.5 and 6.6 are identical with the exception of their exponents. Specifically, the beta distribution has exponents of $a - 1$ and $b - 1$, whereas the binomial distribution has corresponding exponents of y (i.e., the number of successes) and $N - y$ (i.e., the number of nonsuccesses). This similarity is not coincidental, because the binomial and beta densities actually belong to the same distribution family (i.e., the same function describes the shape of the distributions). Specifically, the binomial distribution is a beta distribution in which $a = y + 1$ and $b = N - y + 1$. Researchers frequently adopt priors that belong to the same distribution family as the likelihood function, and this is true of the depression example. These so-called **conjugate distributions** are advantageous because they produce a posterior distribution that also belongs to the same family.

In a previous section, I explained that assigning a number of imaginary data points to the prior determines its influence on the analysis results (the hypothetical sample size is one of the prior distribution's hyperparameters). The equivalence of the beta and the binomial distributions illustrates this point. For example, Researcher A's prior is a beta distribution with $a = 7$ and $b = 40$. A beta distribution with $a = 7$ equates to a binomial distribution with a hypothetical sample of six depressed cases (i.e., $a = y + 1$, so $y = a - 1 = 6$). Similarly, $b = 40$ corresponds to a binomial distribution with 45 imaginary data points (i.e., $b = N - y + 1$, so $N = b - 1 + y = 45$). In contrast, Researcher B's flat prior is a beta distribution with $a = 1$ and $b = 1$. This equates to a binomial distribution with an imaginary sample size of zero (i.e., $y = a - 1 = 0$ and $N = b - 1 + y = 0$). Note that I use the words "hypothetical" and "imaginary" to describe the sample size because the researchers specified their prior distributions before collecting data.

The Posterior Distribution

The final step of a Bayesian analysis is to define the posterior distribution. Ignoring the denominator of Bayes' theorem for the moment, note that Equation 6.4 says that the height of the posterior distribution at each value of π is proportional to the product of the prior times the likelihood. Conceptually, multiplying the likelihood by the prior weights each point on the likelihood function by its prior probability. To illustrate, return to the relative probabilities in Table 6.1. To begin, consider the height of Researcher A's prior distribution at $\pi = 0.05$ and $\pi = 0.10$, the values of which are 0.792 and 6.153, respectively. Multiplying each quantity by its corresponding likelihood gives $0.792 \times 0.106 = 0.084$ and $6.153 \times 0.089 = 0.547$. Visually, 0.084 and 0.547 represent the height of Researcher A's posterior distribution at $\pi = 0.05$ and $\pi = 0.10$, respectively. Consequently, after updating her prior beliefs with information from the data, Researcher A would claim that $\pi = 0.10$ is a more plausible parameter value than $\pi = 0.05$. Turning to Researcher B, the height of his prior distribution was 1.00 at every value of π. Multiplying the prior by the likelihood gives values of $1.00 \times 0.106 = 0.106$ and $1.00 \times .089 = 0.089$. Again, 0.106 and 0.089 represent the height of Researcher B's posterior distribution at $\pi = 0.05$ and $\pi = 0.10$, respectively. Unlike Researcher A, Researcher B would claim that $\pi = 0.05$ is somewhat more plausible than $\pi = 0.10$. However, notice that Researcher B's conclusion is based solely on the data because the shape of his posterior distribution is identical to that of the likelihood function. Again, this is an important consequence of adopting a noninformative prior distribution.

The Role of the Marginal Distribution

Until now, I have ignored the marginal distribution that appears in the denominator of Bayes' theorem. As I explained previously, the **marginal distribution** is a scaling constant that does not influence the shape of the posterior. To understand how the marginal distribution works, consider a simple probability example. Suppose that you wanted to know the probability of flipping a coin three times and getting two heads. Three possible sequences produce this outcome: (1) heads, heads, tails, (2) heads, tails, heads, and (3) tails, heads, heads. By itself, the fact that the three different sequences produce two heads does not provide an accurate gauge of the probability because there is no way of knowing whether three sequences is a large number or a small number. Judging the probability becomes easier after dividing by the total number of possible sequences (there are eight). Now, it becomes clear that 37.5% of the sequences produce two heads. Notice that dividing by the total number of possible outcomes does not change the number of sequences that produce two heads, but it does standardize things in a way that makes the probabilities sum to one.

Using only the numerator of Bayes' theorem is akin to expressing the posterior probabilities on an unstandardized metric (e.g., three sequences produce two heads), and dividing by the marginal distribution standardizes the probabilities (e.g., 37.5% of the sequences produce two heads) such that the area under the posterior distribution sums to one. Conceptually, the marginal distribution works as follows. Suppose that you computed the height of the posterior distribution at every possible value of π by multiplying the prior probabilities by their corresponding likelihood values. Summing these products yields a quantity that is analogous to the total number of possible outcomes from the coin toss example. To illustrate, the bottom row of Table 6.1 sums the product of the prior times the likelihood for integer values of π between 0.05 and 0.20. The value of 3.5338 represents Researcher A's marginal distribution, and 0.9073 is the corresponding value for Researcher B. The Scaled Posterior columns of Table 6.1 divide the posterior probabilities by the appropriate marginal distribution.

Doing so effectively standardizes the height of the posterior distribution such that posterior probabilities sum to one.

In reality, the population proportion can take on an infinite number of values between zero and one, so the example in Table 6.1 is not mathematically accurate. That is, the correct marginal distributions sum the product over every possible value of π, not just integer values between 0.05 and 0.20. With a continuous density function such as the beta distribution, the summation of the prior times the likelihood involves a calculus integral. Nevertheless, whether you think about it as a sum or an integral, the marginal distribution is a constant value that standardizes the height of the posterior distribution such that the total area under the curve sums (i.e., integrates) to one.

6.6 HOW DOES BAYESIAN ESTIMATION APPLY TO MULTIPLE IMPUTATION?

Multiple imputation generates several copies of the data and fills in (i.e., imputes) each copy with different estimates of the missing values. This process uses an iterative algorithm that repeatedly cycles between an imputation step and a posterior step (an I-step and a P-step, respectively). The I-step uses the stochastic regression procedure from Chapter 2 to impute the missing values, and the P-step uses the filled-in data to generate new estimates of the mean vector and the covariance matrix. Virtually every aspect of multiple imputation is rooted in Bayesian methodology, but the ideas from the previous sections are particularly relevant to the P-step because it is essentially a standalone Bayesian analysis that describes the posterior distribution of a mean vector and a covariance matrix.

Generating multiple sets of imputed values requires different estimates of the mean vector and the covariance matrix at each I-step (recall from Chapter 2 that the stochastic regression procedure uses $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ to construct a set of imputation regression equations), and the purpose of the P-step is to generate these parameter values. At each P-step, the iterative algorithm uses the filled-in data from the preceding I-step to define the posterior distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. It then uses Monte Carlo simulation to "draw" new estimates of the mean vector and the covariance matrix from their respective posteriors. The subsequent I-step uses these updated parameter values to construct a new set of regression equations that are slightly different from those at the previous I-step. Repeating the two-step procedure a number of times generates several copies of the data, each of which contains unique estimates of the missing values.

Given the important role that the mean vector and the covariance matrix play in a multiple imputation analysis, the rest of the chapter is devoted to defining the posterior distributions of these parameters. As you will see, the estimation steps remain the same (i.e., specify a prior, estimate the likelihood, define the posterior), but the distribution families are different. Because each P-step uses a filled-in data set, the complete-data procedures described in this chapter are identical to those in a multiple imputation analysis. Finally, it is worth noting that the selection of prior distributions has received considerable attention in the Bayesian literature (e.g., see Kass & Wasserman, 1996). Because the majority of multiple imputation analyses rely on a standard set of noninformative priors, I limit the subsequent discussion to the prior distributions that you are likely to encounter in multiple imputation software packages.

6.7 THE POSTERIOR DISTRIBUTION OF THE MEAN

This section illustrates how to apply Bayesian estimation principles to the mean. I start by applying the three analysis steps to a univariate example and later extend the ideas to multivariate data. To simplify things, I assume that the population variance is known, but this does not affect the underlying logic of the estimation process, nor does it affect the shape of the posterior distribution.

The Prior Distribution

The first step of a Bayesian analysis is to specify a prior distribution. Consistent with the previous depression example, you could specify a prior distribution that assigns a higher weight to mean values that you think are more probable, or you could use a noninformative prior that equally weights every value of the mean. The standard noninformative prior is a flat distribution that assigns an equal weight to every possible value of the mean. The Bayesian literature often refers to this as a **Jeffreys' prior**, after a Bayesian theoretician who proposed a set of principles for developing noninformative priors (Jeffreys, 1946, 1961). Using my previous notation, note that the Jeffreys' prior for the mean is $p(\mu) = 1.00$. In words, the prior states that every possible value of the population mean has the same a priori weight of 1.00. Visually, this prior is identical to the solid line in Figure 6.1.

The Likelihood Function

The second step of a Bayesian analysis is to collect data and use the likelihood function to summarize the data's evidence about different parameter values. Assuming a normal distribution for the population data, the sample likelihood is

$$p(Y|\mu, \sigma^2) = \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5(y_i-\mu)^2/\sigma^2} \right\} \tag{6.7}$$

where braces contain the probability density function for the normal distribution (i.e., the likelihood for a single score), $\prod$ is the multiplication operator, and $p(Y|\mu, \sigma^2)$ is the likelihood of the sample data, given the values of μ and σ^2. (In previous chapters, I used the generic symbol L to denote the likelihood.) Recall from Chapter 3 that substituting a score value and the population parameters into the density function returns the likelihood for an individual score (i.e., the height of the normal curve at y_i), and multiplying the individual likelihood values gives the sample likelihood. Repeating these computations with different values of μ produces a likelihood function that describes the relative probability of the data across a range of population means.

TABLE 6.2. IQ and Job Performance Data

IQ	Job performance
78	9
84	13
84	10
85	8
87	7
91	7
92	9
94	9
94	11
96	7
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	12

To illustrate the likelihood step, consider the IQ scores in Table 6.2. I designed these data to mimic an employee selection scenario in which prospective employees complete an IQ test during their interview and a supervisor subsequently rates their job performance following a 6-month probationary period. These are the same data that I used in Chapter 3 to illustrate maximum likelihood estimation. I used Equation 6.7 to compute the sample likelihood for population mean values between 80 and 120, and Figure 6.4 shows the resulting likelihood function (for simplicity, I fixed σ^2 at its sample estimate of 199.58). The height of the curve is the relative probability that the sample of IQ scores in Table 6.2 originate from a normally distributed population with a mean equal to the value of μ on the horizontal axis and a variance equal to $\sigma^2 = 199.58$. As seen in the figure, the maximum likelihood estimate of the mean is $\hat{\mu} = 100$, which is the same estimate that I derived from the log-likelihood function in Chapter 3.

The Posterior Distribution

The final step of a Bayesian analysis is to define the posterior distribution. The numerator of Bayes' theorem states that the height of the posterior is proportional to the product of the prior times the likelihood. Consistent with the previous depression example, the height of the posterior distribution at any given value of μ is the product of the prior probability and the likelihood. In this situation, obtaining the posterior distribution is simply a matter of multiplying each point on the likelihood function by a value of 1.00. Consequently, the shape

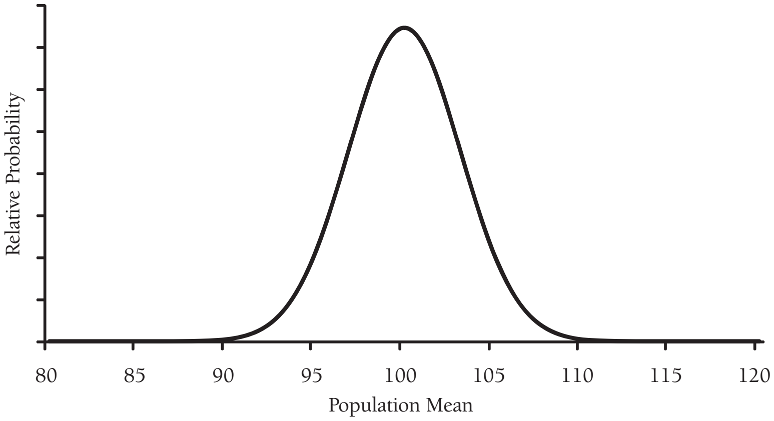

FIGURE 6.4. The likelihood function for the mean. The height of the curve is the relative probability that the IQ scores in Table 6.2 originated from a normally distributed population with a mean equal to the value of μ on the horizontal axis. The maximum of the function (i.e., the maximum likelihood estimate) corresponds with $\mu = 100$, which is the sample mean.

of the posterior distribution is identical to that of the likelihood function in Figure 6.4. More formally, the shape of the posterior distribution is

$$p(\mu \mid Y, \sigma^2) \sim N\left(\hat{\mu}, \frac{\sigma^2}{N}\right) \tag{6.8}$$

where $p(\mu \mid Y, \sigma^2)$ is the posterior distribution, $\sim N$ denotes a normal curve (the ~ symbol means "distributed as"), $\hat{\mu}$ is the sample mean, and σ^2/N is the variance of the posterior. In words, Equation 6.8 says that the posterior distribution is a normal curve that is centered at the sample mean and has a variance of σ^2/N. Notice that the data alone define the shape of the posterior (i.e., the distribution is centered at the maximum likelihood estimate), which is a consequence of adopting a noninformative prior distribution. In addition, the shape of the posterior is identical to the frequentist sampling distribution (e.g., the posterior variance is the square of the usual formula for the standard error of the mean).

The Posterior Distribution of a Mean Vector

A univariate example is useful for understanding the mechanics of Bayesian estimation, but multiple imputation relies on the posterior distribution of a mean vector. Fortunately, the previous ideas readily extend to multivariate data. For example, the standard noninformative prior for a mean vector is a multidimensional flat surface that assigns an equal weight to every combination of mean values. Similarly, the likelihood function is a multivariate, rather than univariate, normal distribution. Finally, the posterior is a multivariate normal distribution that has the same shape as the likelihood function. More formally, the shape of the posterior is

$$p(\boldsymbol{\mu} \mid \mathbf{Y}, \boldsymbol{\Sigma}) \sim MN(\hat{\boldsymbol{\mu}}, N^{-1}\boldsymbol{\Sigma}) \tag{6.9}$$

where $p(\boldsymbol{\mu}|\mathbf{Y}, \boldsymbol{\Sigma})$ is the posterior distribution, ~MN denotes the multivariate normal distribution, $\hat{\boldsymbol{\mu}}$ is the vector of sample means, and $\boldsymbol{\Sigma}$ is the population covariance matrix. Again, the fact that the posterior is centered at the sample means indicates that the prior has no influence on the distribution. Consistent with the univariate example, Equation 6.9 assumes that the population covariance matrix is known, but the equation remains the same when $\hat{\boldsymbol{\Sigma}}$ replaces $\boldsymbol{\Sigma}$.

6.8 THE POSTERIOR DISTRIBUTION OF THE VARIANCE

The covariance matrix plays an important role in a multiple imputation analysis, so it is important to understand its posterior distribution. However, this distribution is more complex than that of a mean vector, and it belongs to a distribution family that is less familiar. Consequently, starting with a univariate example that involves a single variance makes it easier to understand how Bayesian estimation applies to a covariance matrix. As you will see, the ideas in this section readily generalize to a full covariance matrix. For simplicity, I temporarily assume that the population mean is known, but I later describe how the posterior distribution changes when the mean is also a random variable.

The Likelihood Function

The first step of a Bayesian analysis is to specify a prior distribution. Bayesian texts recommend a noninformative prior distribution that looks somewhat different from the flat prior described in previous sections. This new prior will make more sense if you first understand the shape of the likelihood function; I will therefore present things out of order in this section, beginning with the likelihood. Reconsider the normal likelihood in Equation 6.7. Multiplying the collection of bracketed terms by itself N times gives the sample likelihood. After performing this operation, the sample likelihood becomes

$$p(Y|\mu, \sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{N}{2}}} e^{\frac{-.5}{\sigma^2}\Sigma(y_i-\mu)^2} \tag{6.10}$$

The right-most term of Equation 6.10 is the sum of the squared deviations around the population mean. Thus, the likelihood further reduces to

$$p(Y|\mu, \sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{N}{2}}} e^{-.5\left(\frac{SS}{\sigma^2}\right)} \tag{6.11}$$

where SS is the sum of squares. The "proportional to" symbol (i.e., $\propto$) indicates that I omitted the scaling constant (i.e., 2π) from the equation.

Equation 6.11 is useful because it shows how the relative probability of the data (i.e., the sum of squares) varies across different values of the population variance. To illustrate, reconsider the IQ scores in Table 6.2. Assuming a population mean of $\mu_{IQ} = 100$ yields a sum

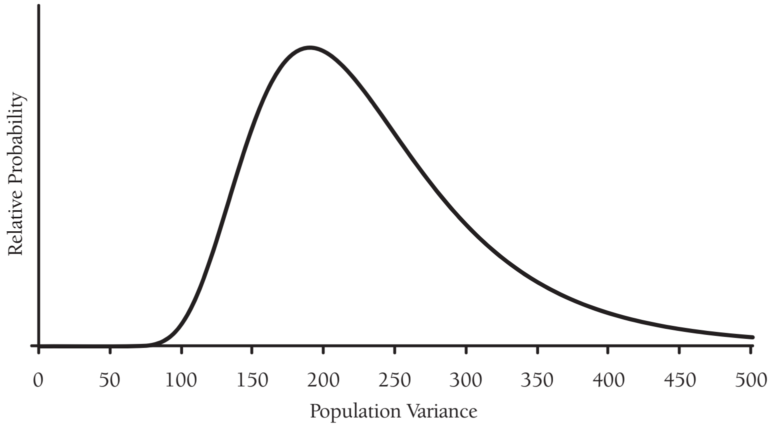

FIGURE 6.5. The likelihood function for the variance. The height of the likelihood function is the relative probability that a sample variance of 189.60 (the variance of the IQ data in Table 6.2) originated from a normally distributed population with a variance equal to the value of σ^2 on the horizontal axis. The likelihood function for the variance belongs to the family of inverse chi-square distributions.

of squares value of $SS = 3792$. I used Equation 6.11 to compute the sample likelihood across a range of population variances, and Figure 6.5 shows the resulting likelihood function. The likelihood function is a positively skewed distribution, but it works in the same manner as before. Specifically, the height of the curve is the relative probability of the data, given the population variance on the horizontal axis. Visually, the maximum of the likelihood function corresponds to a population variance that is slightly less than 200. You may recall from Chapter 3 that the maximum likelihood estimate of the IQ variance was $\hat{\sigma}^2_{IQ} = 189.60$, so Figure 6.5 agrees with this previous analysis.

The likelihood function in Figure 6.5 is an **inverse chi-square distribution**. More accurately, the likelihood is a scaled inverse chi-square distribution, but I simply refer to it as an inverse chi-square throughout the remainder of the chapter. Using generic notation, note that the shape of an inverse chi-square distribution with ν degrees of freedom is

$$\text{Inv-}\chi^2 \propto \frac{1}{x^{\frac{\nu}{2}+1}} e^{-.5\left(\frac{S}{x}\right)} \tag{6.12}$$

where x is a variable, and S is a scale parameter that dictates the spread of the distribution (e.g., larger values of S produce a wider distribution). As before, the "proportional to" symbol (i.e., $\propto$) denotes an omitted scaling constant. Like the chi-square distribution, the inverse chi-square is a family of distributions where the exact shape of the curve is determined by the degrees of freedom (and in the case of a scaled inverse chi-square, the scale parameter).

Relabeling the terms in Equation 6.12 better illustrates the linkage between the likelihood and the inverse chi-square distribution. Specifically, replacing x with σ^2, ν with N, and S with SS gives

$$\text{Inv-}\chi^2 \propto \frac{1}{(\sigma^2)^{\frac{N}{2}+1}} e^{-.5\left(\frac{SS}{\sigma^2}\right)} \tag{6.13}$$

Notice that Equation 6.13 is nearly identical to the likelihood, but σ^2 has an exponent of $(N/2) + 1$ rather than $N/2$. This disparity reflects a difference of two degrees of freedom, so the likelihood is actually an inverse chi-square distribution with $\nu = N - 2$ degrees of freedom.

The Prior Distribution

Having gained some familiarity with the inverse chi-square distribution, I now return to the first step of a Bayesian analysis, which is to specify a prior distribution. Researchers frequently adopt conjugate priors that belong to the same distribution family as the likelihood, so the inverse chi-square distribution is a reasonable prior for σ^2. However, using the inverse chi-square as a prior distribution requires a sum of squares value and an imaginary sample size (i.e., the hyperparameters). Substituting N = 0 and SS = 0 into Equation 6.13 is akin to saying that you have no prior information about the variance. Doing so yields the Jeffreys' prior as follows:

$$p(\sigma^2) \propto \frac{1}{\sigma^2} \tag{6.14}$$

Equation 6.14 is different from the Jeffreys' prior for the mean because it assigns relative probabilities that increase as the population variance approaches zero. To illustrate, Figure 6.6 shows a graphic of the prior distribution, where the height of the curve represents the a priori relative probability for a particular value of σ^2.

The Posterior Distribution

Having established the prior distribution and the likelihood function, the third step of a Bayesian analysis is to define the posterior distribution. As before, the posterior is propor-

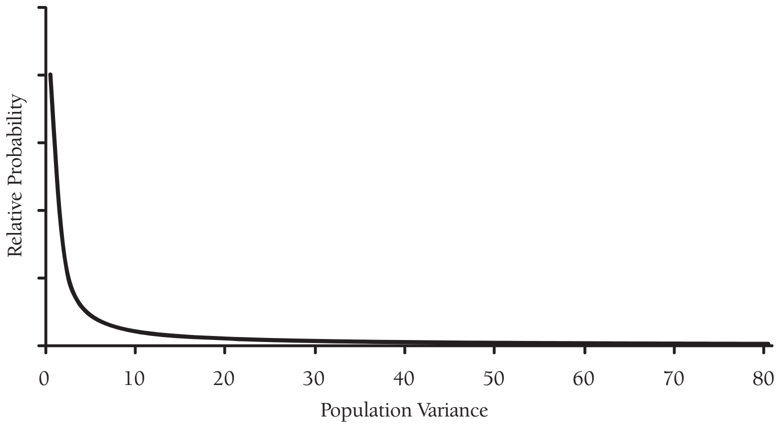

FIGURE 6.6. The Jeffreys' prior for the variance. The height of the curve represents each parameter value's a priori weight. Unlike the Jeffreys' prior for the mean, the prior probabilities increase as the population variance approaches zero.

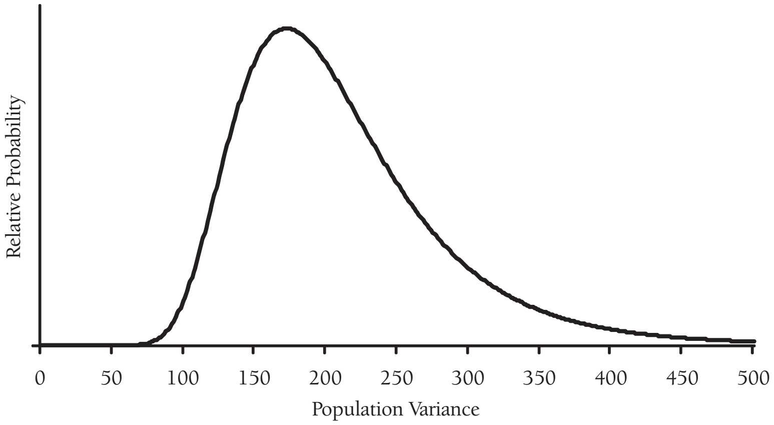

FIGURE 6.7. The posterior distribution of the variance. The posterior is very similar to the likelihood in Figure 6.5, but its left tail is slightly thicker than that of the likelihood. This subtle difference results from using a noninformative prior distribution that assigns higher weights to lower values of σ^2. The posterior distribution belongs to the family of inverse chi-square distributions.

tional to the prior times the likelihood, so the posterior distribution for the variance is as follows:

$$p(\sigma^2 | Y, \mu) \propto \frac{1}{\sigma^2} \times \frac{1}{(\sigma^2)^{\frac{N}{2}}} e^{-.5\left(\frac{SS}{\sigma^2}\right)} = \frac{1}{(\sigma^2)^{\frac{N}{2}+1}} e^{-.5\left(\frac{SS}{\sigma^2}\right)} \tag{6.15}$$

Notice that the posterior distribution is an inverse chi-square distribution with N degrees of freedom and is identical to Equation 6.12. Substituting $SS = 3792$ into Equation 6.15 yields the posterior distribution in Figure 6.7. The effect is subtle, but you can see that left tail of the posterior distribution is slightly thicker than that of the likelihood function, which follows from the fact that the prior assigns higher weights to lower values of the population variance.

Estimation with an Unknown Mean

Throughout this section, I have effectively assumed that the population mean is known. Treating the mean as an unknown random variable changes the shape of the posterior in a way that is analogous to using the sample, rather than the population, formula to compute the variance. Bayesian texts give the mathematical details behind this change (e.g., see Gelman et al., 1995, pp. 67–68), but the result is a marginal posterior distribution with $N - 1$ degrees of freedom (i.e., the exponent of σ^2 changes from $(N/2) + 1$ to $(N+1)/2$). More formally, the shape of the posterior distribution is

$$p(\sigma^2 | \hat{\mu}, Y) \sim \text{Inv-}\chi^2(N - 1, SS) \tag{6.16}$$

where $p(\sigma^2 | Y, \mu)$ is the posterior distribution, ~Inv-χ^2 denotes an inverse chi-square distribution, $N - 1$ is the degrees of freedom, and SS is the sum of squares. The degrees of freedom and sum of squares values are known as the location and scale parameters, respectively, because they determine the expected value and the spread of the posterior distribution (the mean and the variance play a similar role in defining the posterior distribution of the mean). As an aside, the sampling distribution of the variance is also an inverse chi-square distribution with $N - 1$ degrees of freedom; thus, adopting the Jeffreys' prior in Equation 6.14 brings the Bayesian and frequentist paradigms into alignment.

6.9 THE POSTERIOR DISTRIBUTION OF A COVARIANCE MATRIX

This section extends Bayesian estimation to an entire covariance matrix. The basic procedure is similar to estimating a variance, and the distributions are multivariate extensions of the inverse chi-square. By now, you are probably familiar with the three steps of a Bayesian analysis, so I give an abbreviated outline of the process. Consistent with the previous section, I present things out of order, beginning with the likelihood. For simplicity, I temporarily assume that the population means are known, but this does not affect the logic of the estimation process.

The Likelihood Function

Equation 6.11 describes how the likelihood of the sample data varies across different values of the population variance. The corresponding likelihood function for a covariance matrix is

$$p(\mathbf{Y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-N/2} e^{-.5(\mathrm{tr}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}])} \tag{6.17}$$

where $\boldsymbol{\mu}$ is the population mean vector, $\boldsymbol{\Sigma}$ is the population covariance matrix, and $\boldsymbol{\Lambda}$ is the sum of squares and cross products matrix. Equation 6.17 replaces the terms in Equation 6.11 with their matrix analogs, but the likelihood still gives the relative probability of the data (in this case, the sum of squares and cross products matrix represents the data) across different values of the population parameters. To illustrate, Figure 6.8 shows the likelihood surface for a bivariate covariance matrix. I based the likelihood on a sample of 20 cases that I generated from a multivariate normal distribution with means of zero, variances equal to three, and a covariance equal to zero. Notice that the likelihood function is now a three-dimensional positively skewed distribution, but its shape resembles that of the univariate likelihood function in Figure 6.5. Consistent with its univariate counterpart, the height of the likelihood surface at any given point is the relative probability of the data, given the combination of population variances on the horizontal and depth axes.

The likelihood function in Figure 6.8 is a member of the **inverse Wishart distribution** family. The inverse Wishart density function is

$$W^{-1} \propto |\boldsymbol{\Sigma}|^{-(\nu+k+1)/2} e^{-.5(\mathrm{tr}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}])} \tag{6.18}$$

where W^{-1} denotes the inverse Wishart distribution, ν is the degrees of freedom, $\boldsymbol{\Lambda}$ is the sum of squares and cross products matrix, $\boldsymbol{\Sigma}$ is the population covariance matrix, and k is the number of variables. As before, the "proportional to" symbol (i.e., $\propto$) indicates that I

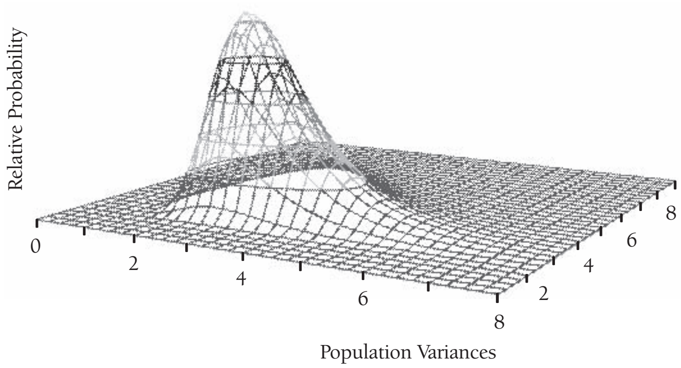

FIGURE 6.8. The likelihood surface for a bivariate covariance matrix. This likelihood is based on a sample of 20 cases from a multivariate normal population with means of zero, variances equal to three, and a covariance equal to zero. The likelihood surface is a three-dimensional positively skewed distribution, but its shape resembles that of the univariate likelihood in Figure 6.5. The height of the likelihood surface at any given point quantifies the relative probability of the sample covariance matrix, given the population variances on the horizontal and depth axes. The likelihood function belongs to the family of inverse Wishart distributions.

excluded a scaling constant from the equation. Notice that the likelihood function and the inverse Wishart distribution are nearly identical, but have different exponents. This is not coincidental, because the likelihood function is an inverse Wishart distribution where ν equals $N - k - 1$. Finally, note that Equation 6.18 reduces to the inverse chi-square distribution in Equation 6.13 when $k = 1$.

The Prior Distribution

Having gained some familiarity with the inverse Wishart distribution, I now return to the first step of a Bayesian analysis, which is to specify a prior distribution. Researchers often choose conjugate priors that belong to the same distribution family as the likelihood, so the inverse Wishart is a reasonable prior distribution for the covariance matrix. Substituting $\nu = 0$ (i.e., zero imaginary data points) and $\mathbf{\Lambda} = 0$ into Equation 6.18 is akin to saying that you have no prior information about the population covariance matrix. Doing so yields the multivariate version of the Jeffreys' prior.

$$p(\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-\frac{k+1}{2}} \tag{6.19}$$

The determinant $|\mathbf{\Sigma}|$ is a scalar value that quantifies the total variation in the population covariance matrix. Because the value of the determinant decreases as variability decreases, the prior probabilities increase as the elements in the population covariance matrix approach zero. This was also true of the Jeffreys' prior for the variance, and Equation 6.19 reduces to Equation 6.14 when $k = 1$.

The Posterior Distribution

The final step of a Bayesian analysis is to define the posterior distribution. Consistent with the previous examples, the height of the posterior distribution is proportional to the product of the prior distribution times the likelihood. Multiplying the prior and the likelihood yields an inverse Wishart distribution with N degrees of freedom. This distribution is identical to Equation 6.18 but replaces ν with N. Like its univariate counterpart, the posterior distribution changes slightly when the means are unknown and becomes an inverse Wishart distribution with $N - 1$ degrees of freedom. More formally, the posterior is

$$p(\mathbf{\Sigma} \mid \hat{\boldsymbol{\mu}}, \mathbf{Y}) \sim W^{-1}(N - 1, \hat{\mathbf{\Lambda}}) \tag{6.20}$$

where $p(\mathbf{\Sigma} \mid \hat{\boldsymbol{\mu}}, \mathbf{Y})$ is the posterior distribution, W^{-1} denotes the inverse Wishart distribution, $N - 1$ is the degrees of freedom, and $\hat{\mathbf{\Lambda}}$ is the sample sum of squares and cross products matrix. In words, Equation 6.20 says that the posterior distribution of a covariance matrix is an inverse Wishart distribution with $N - 1$ degrees of freedom and scale parameter equal to the sum of squares and cross products matrix. The degrees of freedom and sum of squares and cross products matrix determine the expected value and the spread of the distribution, respectively. Importantly, the data (i.e., the sample size and $\mathbf{\Lambda}$) define the shape of the posterior, and the prior effectively plays no role. This has been a consistent theme throughout this chapter and is a result of adopting a noninformative prior distribution. The sampling distribution of $\hat{\mathbf{\Sigma}}$ is also an inverse Wishart distribution with $N - 1$ degrees of freedom, so the Jeffreys' prior in Equation 6.19 brings the Bayesian and frequentist paradigms into alignment.

6.10 SUMMARY

Chapter 7 introduces a second "modern" missing technique, multiple imputation. Rubin (1987) developed multiple imputation within the Bayesian framework, so understanding the nuances of imputation requires a basic working knowledge of Bayesian statistics. The goal of this chapter was to provide a user-friendly account of Bayesian statistics, while still providing interested readers with the technical information necessary to understand the seminal missing data literature (e.g., Little & Rubin, 2002; Rubin, 1987; Schafer, 1997).

Understanding Bayesian statistics requires a shift in thinking about the population parameter. Unlike the frequentist paradigm, Bayesian methodology defines a parameter as a random variable that has a distribution. An important analysis goal is to describe this distribution's shape, and doing so requires three steps. The first step is to specify a prior distribution that describes your subjective beliefs about the relative probability of different parameter values before collecting data. In general, you can specify an informative prior that assigns a higher weight to parameter values that you feel are more probable, or you can specify a noninformative prior that uniformly weights different values—multiple imputation analyses generally use the latter approach. The second step of a Bayesian analysis is to use a likelihood function to summarize the data's evidence about different parameter values. The final step of a Bayesian analysis is to define the parameter's posterior distribution. Multiplying the likelihood by the prior distribution adjusts the height of the likelihood function up or down according to the magnitude of the prior probabilities and yields a new composite distribution that describes the relative probability of different parameter values.

Because the mean vector and the covariance matrix play an important role in a multiple-imputation analysis, a key goal of this chapter was to define the posterior distributions of these parameters. The posterior distribution of a mean vector is a multivariate normal distribution, whereas the posterior distribution of a covariance matrix is an inverse Wishart distribution. The majority of multiple imputation analyses rely on a standard set of noninformative prior distributions (i.e., so-called Jeffreys' priors). Adopting a Jeffreys' prior effectively eliminates the influence of the prior distribution and yields a posterior distribution that is defined solely by the data. The Jeffreys' priors also bring the Bayesian and the frequentist paradigms into alignment because the posterior distributions of the mean vector and the covariance matrix are identical to the frequentist sampling distributions.

The next chapter introduces multiple imputation. Multiple imputation is actually a broad term that encompasses a collection of different techniques, but I focus on a data augmentation algorithm that assumes a multivariate normal distribution (Schafer, 1997; Tanner & Wong, 1987). Data augmentation is an iterative algorithm that repeatedly cycles between an I-step and a P-step (i.e., an imputation and a posterior step, respectively). The I-step uses the stochastic regression procedure from Chapter 2 to impute the missing values, and the P-step defines the shape of the posterior distributions and uses Monte Carlo simulation to "draw" new estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from their respective posteriors. Repeating this two-step procedure a number of times generates several copies of the data, each of which contains unique estimates of the missing values. The posterior step is essentially a standalone Bayesian analysis of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, so the ideas in this chapter play an important role throughout Chapter 7.

6.11 RECOMMENDED READINGS

Bolstad, W. M. (2007). *Introduction to Bayesian statistics* (2nd ed.). New York: Wiley.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall.

Lee, M. D., & Wagenmakers, E-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.

Pruzek, R. M. (1997). An introduction to Bayesian inference and its applications. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 287–318). Mahwah, NJ: Erlbaum.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling: An Interdisciplinary Journal*, *11*, 424–451.