CHAPTER 24

BAYESIAN MODELING FOR PSYCHOLOGISTS: AN APPLIED APPROACH

Fred M. Feinberg and Richard Gonzalez

Bayesian methods offer new insight into standard statistical models and provide novel solutions to problems common in psychological research, such as missing data. Appeals for Bayesian methods are often made from a dogmatic, theory-based standpoint concerning the philosophical underpinnings of statistical inference, the role of prior beliefs, claims about how one should update belief given new information, and foundational issues, such as the admissibility of a statistical decision. Although such a rhetorical approach is academically rigorous, it usually is not the kind of argument a practicing researcher wants to read about. Researchers care about analyzing their data in a rigorous manner that leads to clear, defensible conclusions. In this chapter, we address the reader who wants to learn something about what all the Bayesian fuss is about and whether the Bayesian approach offers useful tools to incorporate into one's data analytic toolbox. We hope this chapter prompts readers to learn more about what Bayesian statistical ideas have to offer in standard data analytic situations. Throughout the chapter, we highlight important details of the Bayesian approach; how it differs from the frequentist approach typically used in psychological research; and most important, where it offers advantages over the methods most commonly used by academic researchers in psychology and cognate disciplines.

SOME GENTLE PRELIMINARIES

Practicing research psychologists wish to understand and explain a variety of behaviors in humans and animals. Statistical methods and reasoning sharpen insight into experimental design and avoid the potential pitfalls of lay examination of data patterns. Deserving special mention is a point often missed in substantively focused studies: The purpose of statistical inference is to replace intuitions based on a mass of data with those achievable from examination of *parameters*. Except in nonparametric settings relatively rare in psychological and other social science research, understanding one's data relies critically on choosing an appropriate statistical model and both estimating and examining the distributions of its parameters. By this we mean the so-called *marginal distribution*—that is, everything we can say about a parameter once our data have been accounted for.

Too often, researchers shoehorn their hypotheses, which often concern individual-level behavior, into the straightjacket mandated by classical statistical methods. This approach typically requires large numbers of respondents for the central limit theorem to kick in, presumes equal variances in analysis of variance (ANOVA) designs, makes various untested assumptions about (lack of) correlation in errors and variables, requires balanced designs, and so on. Each of these requirements is necessary because the commonly used classical statistical tests do not achieve "nice" forms when their assumptions are violated. Imagine instead a world in which researchers can simply collect a data set and let the chosen statistical model summarize everything of interest it contains; the only assumptions one makes

We presented an earlier version of this chapter at a tutorial workshop on Bayesian techniques at the University of Michigan in March 2007. We thank the College of Literature, Science, and Arts and the Statistics Department of the University of Michigan for its financial support.

DOI: 10.1037/13620-024
APA Handbook of Research Methods in Psychology: Vol. 2. Research Designs, H. Cooper (Editor-in-Chief)

concern the underlying model generating the data and not aspects of the data set itself (e.g., balance, lack of error correlation, and so on); missing values do not mean throwing out a subject's data entirely; individuals can differ in their parameters; and covariates, like age and gender, can be used to describe how and why parameters differ across respondents.

Classical methods, such as nonparametric tests, can sometimes be used in the sorts of situations in which standard assumptions (like underlying normality) are known (or suspected) to be violated. But they typically come at a substantial cost in power: the ability to detect incorrect hypotheses. Bayesian statistical methods, however, provide a general framework that is adaptable to many different types of data, for the relatively modest—and steadily decreasing over time—price of additional computational effort. As we emphasize throughout this chapter, Bayesian methods dramatically expand a researcher's ability to work with real data sets and explain what they have to tell us. Bayesian methods do this by yielding marginal distributions for all parameters of interest, not merely summary measures like means and variances that are only asymptotically valid, and with many fewer presumptions about model forms and large-sample properties of estimators. It is for these reasons that we advocate their increased adoption by the psychological community. In this chapter, we take a first relatively nontechnical step in explaining how this might come about and what Bayesian methods might offer the practicing psychologist.

Many treatments of Bayesian statistics that have been written for (or by) psychologists have focused on the more philosophical issues. Some of these reviews have been made in the context of what is called the *null hypothesis debate*. Practicing research psychologists have become dissatisfied with conventional hypothesis testing and the mental gymnastics that one must undertake in most 1st-year psychology statistics courses to understand its underlying concepts. Examples include how one interprets the usual *p* value, as reflecting the probability of observing some sample statistic under the null hypothesis, or the classical interpretation of a confidence interval as the frequency of intervals that contain the true population parameter value. It is in this context that Bayesian techniques are usually discussed as an alternative way of thinking about intervals, what a Bayesian calls *credible intervals*, and as a similarly different way to think about hypothesis testing, one that avoids many of the conceptual difficulties of the traditional *p* value. It has been argued that the Bayesian approach provides an alternative methodology and philosophical foundation for hypothesis testing (e.g., Jaynes, 1986).

A simple way of conceptualizing the distinction between the two approaches is about how one views uncertainty. A classical statistician views uncertainty as residing in the data one happens to observe: One needs to think about all the other observations that *could* have been made, under the hypothesized model, and base one's statistical test on the resulting distribution, which often achieves a "nice" form (e.g., one that can be looked up in a table). An example of this kind of logic is seen in the Fisherian interpretation of the *p* value (the probability of possible results that are "more extreme" than the observed result) and in some standard tests like the Fisher exact test for contingency tables, which uses the hypergeometric distribution to compute the probability of all contingency tables that are "more extreme" than the one that was actually observed.

The Bayesian approach places uncertainty not in the observations but rather in one's lack of knowledge. For a Bayesian, the observed data are not uncertain—you observed what you observed. But uncertainty has to be addressed somewhere in the analysis. A Bayesian places the uncertainty in our lack of knowledge about *parameters* and operationalizes that lack of knowledge in terms of a (joint) probability distribution over all unknown quantities, that is, parameters. Before any data are observed, the Bayesian summarizes everything known about the model's parameters in just such a distribution, called the *prior distribution*. This can include information from previously conducted studies, common-sense reasoning (e.g., gaining an inch in height will, all else equal, entail an upswing in weight), or even seemingly inviolable facts about parameters (e.g., variances cannot be negative). The prior distribution is then combined with (often called *updated by*) the likelihood, which is common from the usual frequentist analysis, to yield the

posterior distribution. As we will see, literally everything researchers might wish to say about their data—estimation, testing, prediction, and so on—can be extracted, in a natural and direct way, from this posterior. In a sense, it replaces the entire canon of specialized test procedures so laboriously mastered in introductory statistics courses with one simple conceptual object.

In the next section, we refine and illustrate some of these issues, using elementary examples common to statistics texts of both the frequentist and Bayesian varieties. We also provide references to some presently available software and a few comprehensive, book-length treatments of Bayesian statistical methods. Throughout, we eschew formulas and other mainstays of *rigor* for a more user-oriented discussion, one especially geared to the practicing researcher in psychology.

THE NITTY-GRITTY OF THE BAYESIAN APPROACH

Estimating a Proportion

We begin with a relatively simple example, one common throughout statistical inference, in psychology and elsewhere: estimating the proportion of times a particular event occurs. To provide a specific context, consider a dependent variable that codes whether a couple has divorced within their first 20 years of marriage. The data set includes 10 couples, six of which were divorced within the 20-year window. Of course, any beginning student knows that this sample can be used to estimate the divorce rate: simply divide the number of divorces by the total number of couples, 6/10 = 0.6. But how do we know that is the best estimate of the true divorce rate in the population? How do we assess the uncertainty of this estimate?

To handle such questions within the classical framework, one reverts to the likelihood principle (i.e., "all the information in our sample is contained in the likelihood function"), makes an assumption about the independence of those 10 observations, and assumes the binomial model for the observed outcomes. To derive the usual maximum likelihood estimator for the proportion, we take the first derivative of the likelihood, set it to zero, and solve for any unknown parameters, of which in our present example there is only one. Some of the computations in maximum likelihood estimation are simpler if one works with the logarithm of the likelihood—which, as a monotonic transformation, leaves the maximum intact—thus converting products to sums. In our exposition, we focus primarily on the likelihood itself because that is more convenient for Bayesian derivations, and point out when the log likelihood is used instead.

It is a common quip that the likelihood is the only thing about which both Bayesians and frequentists agree, and it is true that the likelihood plays a critical role in both accounts of statistical inference. In simple language, the likelihood function tells us how likely the parameters are to take one set of values, compared with any other. It is not a probability itself (indeed, it can even be greater than 1) but a *relative* statement, so that the likelihood ratio, a common concept in hypothesis testing, is a simple way to assess the comparative degree of appropriateness for any two given sets of parameters. In general, the likelihood is defined by

$$L(\theta \mid Y) = f(Y \mid \theta), \tag{1}$$

where Y represents the observations, θ represents the unknown parameters, and f is some probability density function. It is necessary to assume a distribution f, such as the binomial or normal (perhaps the two most common statistical models), to use maximum likelihood as the basis for parameter estimation. So, even within this classical approach, an important distributional assumption is made at the outset to estimate parameters of interest (e.g., a single population proportion in the binomial case, or the population mean and variance in the normal). It is therefore critical to conceptualize parameters as belonging to a specific model; it is the form of the model's likelihood function that allows the parameters to be estimated, regardless of whether the estimation is classical or Bayesian in nature.

To return to our sample problem, the likelihood for binomial data is given by

$$L(\pi \mid Y) = \binom{N}{Y} \pi^{Y} (1-\pi)^{N-Y}, \tag{2}$$

where π is the population proportion that is being estimated (in the binomial case, the unknown parameter θ is traditionally denoted as π). The number of trials (N) and the number of successes Y (which are the observations) are held fixed, and one searches for values of π that maximize the value of Equation 2. We say *values* because many likelihood functions can have multiple *local maxima*, only one of which is the true global maximum, that is, the single best choice of parameter(s). It is for this reason that maximum likelihood is conceptualized in terms of a search for unknown parameter(s), which in practice is a serious limitation for the classical approach, because multivariate optimization can be exceptionally computationally intensive.

Although the likelihood of Equation 2 may look just like an elementary statement about the probability of observing a particular set of data, in actuality, the inference is done in the other direction; that is, we infer parameter π given the data Y. In the classical estimation approach, the standard error of the parameter emerges by taking the expected value of the Hessian (the matrix of second derivatives) of the log likelihood. The logic justifying the use of the Hessian for this purpose involves imposing assumptions—most notably that the curvature of the log likelihood can be approximated by a multivariate Taylor series, up through its quadratic term. This is the underlying rationale for the typical approach taken by psychologists, computing a point estimate and constructing a confidence interval around that estimate. The classical approach focuses only on the maximum value of the likelihood (the point estimate) and approximates uncertainty (via the Hessian); all other details of parameter estimation are discarded in the classical approach. In this way, the classical statistician is forced to rely on a number of asymptotic (i.e., large sample) assumptions, without any practical way to verify them. It is only when these assumptions hold that the usual properties of estimators, like normality, can be shown to hold. When a problem comes along for which none of the typical distributions (z, t, F, chi-square, etc.) are provable asymptotic approximations, defensible inferences from the data become difficult, if not impossible. As we shall see, the Bayesian is not hampered by this restriction because Bayesian analysis yields the *actual* distributions of any desired set of parameters (or functions of them) and rarely needs to call on the common distributions drilled into every beginning student of statistical inference.

The Bayesian approach also uses the likelihood (Equation 1 in general, or Equation 2 for our binomial example) but differs in how it is used. Although the classical statistician *maximizes* the likelihood by choosing the best parameter values θ, the Bayesian instead converts the problem into a statement about the (posterior) distribution θ. To keep notation simple and not have to keep track of different density functions, we use so-called bracket notation, which has become the standard way to represent useful properties and rigorous derivations in Bayesian analysis. As mentioned earlier, a key Bayesian property is that the posterior distribution is proportional to the product of the likelihood and prior distribution; this will be denoted as

$$[\theta|y] \propto [y|\theta][\theta]. \tag{3}$$

The prior distribution $[\theta]$reflects what we know (or do not know) about the parameters θ *before* consulting the data; the posterior distribution $[\theta|y]$ reflects what we know about the parameters θ *after* combining both the observed data and the information contained in the prior. Bayesians jump freely between talking about probabilities and talking about distributions when referring to priors and posteriors. We will also follow the convention of the field and speak only about proportionality ($\propto$) because this is all that is required for standard Bayesian inference techniques to be applied, a topic we return to later.

The change in reference turns out to be the key property of the Bayesian approach: Rather than work only with the likelihood $[y|\theta]$, as in the classical approach, Bayesians work with the posterior distribution $[\theta|y]$ (a quantity classical statisticians seek to make inferences about but that work in reverse direction with the likelihood). Under this approach, the posterior tells us literally everything we can know about the parameters θ once the data y are observed. The Bayesian merely has to *explore* the posterior and use it for inference. This is simple, at least in theory, but we need to explain what we mean by exploring the posterior.

We must also stress that it is not the case that Bayesians have the extra step of imposing a prior distribution whereas classical statisticians do not. There is a sense in which, when viewed through Bayesian eyes, classical statistics presumes that all values of a parameter are equally likely, what is called a *noninformative* prior in the Bayesian context. A Bayesian with a noninformative prior works with the same functional information as a classical statistician and has an analogous approach to such key issues as the interval constructed around a parameter estimate. The reason is that, with a suitably chosen noninformative prior, the Bayesian posterior is functionally the same as the classical likelihood. So, as long as the Bayesian acts on the posterior in the same way as the classical statistician (e.g., computes the maximum, also called the *mode*), then the two approaches yield identical results. Bayesians, however, provide a different, and some would say more nuanced, description of uncertainty and avoid some of the difficulties that plague classical analyses, such as problems with missing data and unbalanced designs in multilevel models.

The Bayesian framework also provides a language that is more natural to empirical researchers. For example, the Bayesian tradition does not promote the mind-twisting language of how to interpret a confidence interval (i.e., the percentage of such intervals that contain the true population value) and can more directly talk about an interval as representing a 95% degree of confidence (*credibility*) for the value of the unknown parameter. In our experience, students first encountering statistics are put off by counterfactual notions concerning what might happen if similar data were collected many times under identical circumstances. Rather, they ask conceptually direct questions about what can be said using *this* data set, as it is. Bayesian inference refers to the data one has, not the data one might obtain were the world to replay itself multiple times.

Conjugate Priors

One way to simplify calculating and sampling from the posterior—the main practical challenges in most Bayesian analyses—is by careful selection of the prior distribution. There are well-known prior-likelihood pairs, as on the right-hand side of Equation 3, that yield posteriors with the same *form* as the prior (i.e., the prior and posterior fall into the same distributional family, differing only in their parameters). For example, for binomial data, a beta prior distribution (i.e., beta-binomial pair) leads to a beta posterior distribution. Such *conjugate* priors make the overall Bayesian analysis easier to work with, both in terms of derivation and computation (see Box & Tiao, 1973/1992). For example, the beta distribution has two parameters, α and β, in its "functional" portion (i.e., leaving out constants that allow it to integrate to 1), $x^{\alpha-1}(1-x)^{\beta-1}$. Different values of α and β lead to different prior distributions over the unknown parameter π of the binomial distribution, making some values more likely than others *before recourse to the actual data*. For instance, the parameter pair $\alpha = 1$ and $\beta = 1$ produces a beta distribution that is uniform over [0,1], meaning that the prior presumes all values of the unknown binomial parameter π are equally likely; $\alpha = 2$ and $\beta = 2$ makes values of π near one half somewhat more likely; and $\alpha = 101$ and $\beta = 201$ makes values of π near one third quite a bit more likely (all these statements can be verified by simply graphing $x^{\alpha-1}(1-x)^{\beta-1}$ for the values in question). One then conducts an empirical study and observes Y successes out of N trials, such as in the example of six divorces in 10 couples (equivalently, six divorces and four nondivorces). The posterior distribution, when using a conjugate prior (i.e., the prior is beta and the likelihood is binomial), will also be a beta distribution, but the posterior parameters characterizing the posterior distribution are $\alpha + Y$ and $\beta + N - Y$, respectively. In other words, the posterior beta distribution has parameters that consist of both the prior parameter and the data (e.g., the first parameter of the posterior distribution is the sum of the prior parameter α and the observed number of divorces Y, and the second parameter is the sum of the prior parameter β and the number of nondivorces, $N - Y$). So, for our ongoing example, a uniform prior over the [0,1] interval (all proportions are equally likely), leads to a posterior that is a beta distribution with parameters 7 and 5 (i.e., $\alpha = \beta = 1$, $Y = 6$, $N = 10$). For reasons that can now be clearly seen, this process is often referred to as *updating* the

prior, using the data to obtain the posterior from which all Bayesian inference follows.

The mode (i.e., the most likely value, where the density function reaches its largest value) of the beta distribution is $(\alpha - 1)/(\alpha + \beta - 2)$. So, the mode of the posterior for our ongoing divorce example using a uniform prior distribution is six tenths (i.e., the posterior has parameters $\alpha = 7$ and $\beta = 5$, so the mode is six tenths), the same value as the maximum likelihood estimator. Other summary measures of the posterior distribution are also possible. The mean of a beta distribution is $\alpha/(\alpha + \beta)$, so a Bayesian could take the estimate of the proportion to be 7/12 = 0.58, rather than the classically derived maximum likelihood estimate of 6/10 = 0.60 (see Figure 24.1). As the sample size gets larger, the impact of the particular choice of the prior becomes less influential. If we had 600 couples divorce out of 1,000, then the posterior mean would be 601/1002 = 0.5998, which is very close to the maximum likelihood estimate of 0.60. The mode remains .6 for the uniform prior distribution.

We note that the classical and Bayesian estimates for the proportion coincide when the Bayesian uses a beta prior distribution with parameters α and β both very close to zero (note that α and β must be positive to give rise to a nondegenerate distribution; a so-called improper prior results if one literally sets both to zero because the functional part of the beta density, $x^{-1}(1-x)^{-1}$, does not have a finite integral over [0, 1]). Such an improper prior corresponds to a noninformative prior over the logit scale—that is, every value of the logit transform of π, or $log[\pi/(1 - \pi)]$, is equally likely—which can be used to provide a Bayesian justification for using logistic regression in the case of data that follow a binomial distribution.

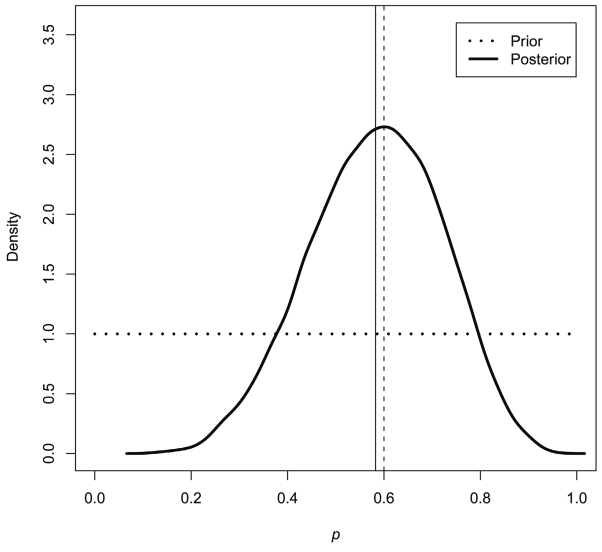

FIGURE 24.1. An example of six divorces out of 10 couples. The solid vertical line is the Bayesian estimate for the unknown proportion given a uniform prior (dotted horizontal line) and posterior distribution (thick solid curve). The theoretical posterior distribution is a beta; the thick solid curve is the estimated posterior distribution from MCMC sampling. The dashed vertical line is the maximum likelihood estimate (0.6).

The prior distribution has other effects on the analysis as well. For example, the prior can define the feasible search space while exploring the posterior distribution. Many statistical models impose restrictions on possible parameter values, such as that variances cannot be negative, which would seem to be an inviolable property that need not be specified, or even checked. Under classical estimation routines, however, negative variances can and do occur, especially in the case of mixture models, as users of programs that estimate such models soon discover. The prior distribution can address these issues by defining the effective feasible region for the unknown variance to be nonnegative (i.e., forcing the prior distribution to have mass only over nonnegative values). A recent example of using the prior to limit the search space of a parameter is the new Bayesian feature in the structural equation modeling program SPSS Amos. The user can specify priors that include, for instance, probability mass over the nonnegative real numbers, thus allowing one to place boundary conditions on the value of variances.

The Whole Distribution and Nothing But the Whole Distribution

The previous subsection sells the Bayesian approach short. Comparing the Bayesian posterior mean to the parameter that emerges from the classical maximum likelihood framework is playing to the strength of the classical approach, which provides the estimate of the mean, and with a little more work and a few additional assumptions, a symmetric standard error (based again on asymptotic assumptions) emerges. The Bayesian approach has much more to offer, however. Instead of providing a point

estimate and a standard error for a given parameter, the Bayesian approach provides its entire posterior distribution. This is available in the form of a sample *drawn* from the posterior, from which any quantity of interest can be computed. In some (all too rare) instances, as in the case of conjugate priors, closed-form solutions are possible, and so the entire posterior distribution takes a known form, and sample draws from the posterior are not needed. But, even when the posterior does not take such a known form—and this is the case in the vast majority of real-world applications—the researcher can, using the posterior, easily compute not only the mean but also the median, the mode, or any quantile one would like (such as the lower or upper 2.5%), even the entire shape of the distribution. Knowledge of the posterior distribution allows one to construct intervals for parameters of interest without making specific, potentially unfounded, assumptions, like symmetry (too often assumed, particularly so for bounded and skewed quantities like variances). Indeed, if the posterior distribution is asymmetric, one can easily construct an interval that accurately reflects this asymmetry.

Furthermore, when the unknown parameter θ is a vector, the posterior distribution becomes multidimensional. The classical approach to a parameter vector θ is to work with point estimates for each parameter separately, calculate a covariance matrix across parameters (via the Hessian), and rely on asymptotic results to make inferences. The Bayesian tackles the entire multivariate posterior distribution head on, however, by taking large samples from it directly. A few decades ago this was rather difficult to do for real-world data sets and models. Modern computers, along with some sampling techniques that emerged from statistical mechanics problems in physics (e.g., Gibbs sampling, simulated annealing), have revolutionized how an analyst can explore a posterior distribution. As mentioned, the solution turns out to be sampling: One takes a sample as large as one needs from the posterior distribution and uses that sample for inference and model comparison.

It is important to distinguish the Bayesian approach to sampling from the posterior from well-known *resampling* procedures, such as the bootstrap or the jacknife. The Bayesian approach produces samples of the unknown parameters, whereas other approaches to estimation or inference that make use of sampling involve sampling either the observations themselves or quantities from a model fitting procedure. For example, to bootstrap slopes in a regression equation, one can either create bootstrap samples of the original data set and compute regression parameters on each of those bootstrap samples (in which case there are as many regressions as bootstrap samples), or one can take bootstrap samples of the residuals after having fit a regression model to the original data (in which case only one regression is estimated). In neither case is the sampling done from the joint posterior distribution for all unknown model quantities, the cornerstone of Bayesian estimation.

Once the multivariate distribution of the parameter vector θ is in hand, one can use it in creative ways to solve some difficult problems. As mentioned, one can compute the mean or median of the posterior distribution. More interesting, one can compute *functions* of the unknown parameters. For example, one common test used in mediation analysis is the product of two unknown parameters: the slope of the predictor to the mediator and the slope of the mediator to the dependent variable (where the latter slope is computed in the context of a regression that also includes the predictor). The prospect of a well-behaved statistical test on the product of two regression slopes is nearly hopeless using frequentist techniques because it cannot be guaranteed to have a *standard* distribution. But the Bayesian perspective provides a well-behaved and reasonable solution. The analyst simply multiplies the samples of the two regression slopes (i.e., multiplies draws from the posterior distributions for both quantities) and instantly has a new posterior distribution for the product. One can then work with that posterior distribution in the usual way, such as compute the mean or median, or the 2.5% and 97.5% quantiles, to construct an interval that can be used to test the hypothesis that the population product is zero. One works with the posterior distribution directly without having to assume symmetry in the sampling distribution (as the classical approach requires, which is suspect in any case because the distribution of a product of random variables is not,

in general, symmetric). Yuan and MacKinnon (2009) in fact provided an introductory account of how to implement this idea for testing mediation in normally distributed data.

It is relatively straightforward to extend this Bayesian approach to mediation to more complicated situations, such as when the mediator or the outcome (or both) involves binary data. In this setting, it is necessary to use a generalized linear model, such as logistic regression, for mediation, and the inference within the classical approach for products of parameters across two such general regressions becomes even more difficult. The Bayesian approach can easily handle mediation models in cases in which predictor, mediator, and outcome are on different scales (such as normally distributed, binary, ordinal, count, or survival data), and it can even be extended into new territories that have not been fully explored within the classical framework (such as a mixture model for mediation in which the analysis partitions the sample into different subgroups exhibiting different mediation patterns). So long as we can sample from the posterior, we can construct any interval or test of interest with little additional effort.

Another example that is relatively simple within the Bayesian approach is the statistical test for a random effect term. Many multilevel modeling programs provide a classical test for the variance for a random effect term against the null value of 0. Unfortunately, the classical test does not apply when testing a parameter at its boundary (i.e., variances are bounded below by 0). So, testing the variance against a null of 0 corresponds to a test that technically does not exist and erroneously produces significant results in all but very small samples. Thus, most tests for the variance of the random effect term that appear in popular programs are, if not overtly incorrect, potentially misleading. Some attempts have been made to address this issue using frequentist methods, but a Bayesian approach handles this problem directly, by yielding the posterior distribution of the variance term under a prior that is properly defined over the feasible range of the variance (a common one being a noninformative prior for the log of the variance). Bayesian testing procedures can compare measures of model fit for a model with a random effect to one without it, akin to the classical likelihood ratio test but valid for testing any set of candidate models against one another, not merely parametrically related (i.e., nested) ones.

Data, Parameters, and Missingness

The shorthand notation of θ to denote the unknown parameters masks the strength of the Bayesian approach. Any and all unknown quantities can be incorporated into the vector θ. For instance, missing data can be construed as unknown parameters and included in θ. The Bayesian practice of estimating the joint distribution enables one to properly capture the effect of missing data on the parameters of interest, such as a mean or regression slope; the overall uncertainty resulting from *all* unknown quantities are jointly modeled. Other unknowns that can enter the vector θ include terms representing random effects and those representing proportions or latent class indicators in mixture models. For each of these features of the Bayesian approach, the entire posterior distribution for all unknowns is estimated: We have not only the point estimate for the missing data but also their posterior distribution, and all other parameters are adjusted for the uncertainty because of the entire *pattern of missingness*. By comparison, the options built into frequentist statistical programs common in psychological analyses—casewise or listwise deletion, or the downright dangerous option to replace missing data by means—appear almost primitive.

Although this chapter lacks space for a full explication of these ideas, one of the major conceptual and computational advantages of the Bayesian approach is its recognition of just two kinds of quantities: Those you know (data) and those you do not know (parameters). Gone are the tedious distinctions between data types, latent variables, limited–censored–truncated, dependent versus independent, missing points or covariates, and the entire menagerie of specialized techniques one must master to deal with them. A Bayesian can simply treat anything not observed as a parameter in the model and, in a rigorous and natural way, numerically integrate over it. So, missing data includes not only literal morsels of unavailable information but also other

unobservables such as latent variables or mixing parameters in a mixture model. Using a technique called *data augmentation* that fills in any missing values (which are treated as parameters) on each pass of the numerical simulator, dramatic simplifications in programming the likelihood are possible. As stated, a full description is well beyond the frame of this chapter. In our view, the ability of Bayesian analysis to seamlessly handle missing data is among its most powerful practical advantages, once researchers properly conceptualize the notion of missingness. We refer the interested reader to the classic texts by Little and Rubin (2002) and Gelman, Carlin, Stern, and Rubin (2004).

Techniques for Sampling From the Posterior Distribution

The idea of sampling from the posterior, and using the sample to compute summary measures such as expected value (means) and the distribution of parameters θ, is the modern contribution of the Bayesian framework. Bayesian computations were extremely difficult before this development of dedicated simulation techniques.

The key innovation in the Bayesian toolbox is the general technique of Markov chain Monte Carlo (MCMC) methods. The basic idea is to sample each unknown parameter in turn (including those that reflect missing data), sequentially cycling through each unknown many times, always conditional on the latest draws for all the others. Under fairly general conditions (which are both technical and satisfied in the vast majority of actual research settings), theorems show that the sampling will reach a *stationary distribution* for the parameters of interest. One of the complexities of Bayesian analysis is that one can only rarely sample from the desired posterior distribution immediately because this would require knowing approximately where it is largest. Instead, one can choose a start point at random, let the simulation go, and, usually within several thousand iterations, a stationary distribution is reached, after which everything produced by the simulator can be used for testing, inference, and forecasting.

Several diagnostic tests are available to identify when such a stationary distribution has been reached. Within the Bayesian framework, the stationary distributions are reached by sampling from the so-called conditional densities (i.e., probability densities for one or a set of parameters given specific values for all the others), but the researcher is interested in—and obtains—samples from the entire joint distribution for all unknown quantities (parameters). Samples from the stationary distribution then serve to estimate parameters and the uncertainty in each as well as assess model fit. The availability of the joint distribution allows for tests that are sometimes difficult, or practically impossible, within the standard framework. For example, if one wants to test the distribution of a product of two unknown parameters (a situation that arises in testing mediation models), it is straightforward to have the product distribution merely by multiplying the samples of the two unknown distributions (Yuan & MacKinnon, 2009). Additionally, it is trivial for the researcher to place a priori constraints on parameters, for example, specifying that the covariance matrix for random effects be diagonal or that specific parameters are uncorrelated. This can be done via the prior or within the sampling scheme, simply by setting any parameter or function of them to a specific value, like zero, and sampling for all the others conditional on the constraints. The analogous procedure in a frequentist analysis can be fantastically difficult, as such constraints can wreck asymptotic normality. But this poses no problems for Bayesians, who need not bother about asymptotics and presumptions of standard distributional forms.

Different methods lie within the MCMC family of algorithms, the dominant ones being Gibbs sampling and Metropolis-Hastings sampling. Loosely put, the former is used when conditional densities take "nice" forms (known distributions relatively easy to sample from), the latter when they do not. (For a good review of these methods, see Tanner 1996.) Bayesian algorithm design is complex and technical, so we cannot provide anything close to a complete description of the subject here. We can, however, readily convey the flavor of what is involved in a nontechnical way. The primary goal of Bayesian analysis is generating a sample from the posterior distribution. This means that the probability that a point (i.e., a set of parameter values) is in

the sample is proportional to the height of the posterior distribution at that point. Or, more usefully, the ratio of the probability of any two points being in the sample is the ratio of the posteriors at those points. This is very close to the foundational insight of the dominant algorithm, Metropolis-Hastings, used in Bayesian analysis: If one is at a point that has already been accepted into the sample, one jumps to another point on the basis of whether its posterior is higher or lower. If it is higher, one jumps; if not, one jumps with probability related to the ratio of the posteriors. (There are some technicalities involving how one generates potential jumps as well, but this would take us far afield.)

This simple algorithm can, in principle, be used to navigate high-dimensional parameter spaces, that is, to estimate statistical models with dozens or even hundreds of parameters. In practice, there are many techniques used to make it efficient, like jumping along one dimension at a time, taking small steps (which make jumping more likely), and using special schemes to choose where to jump. If one can calculate closed-form expressions for particular densities (describing where to jump), it is possible to prove that one always jumps, eliminating a potentially long series of staying put.

When a large number, usually several tens of thousands, of such jumps have been made, one has that many drawn parameter values that can be used for inference. Unfortunately, these draws are often highly autocorrelated. In simple language, this means they do not jump around the distribution randomly, but rather move across it slowly, because where you jump *to* depends on where you jump *from*. In practice, one solves this problem via *thinning*, that is, by discarding all but every 10th, 20th, or 50th draw (the proportion is chosen by the researcher, using various assessment tools). Even with thinning, the researcher will typically have many thousands of points to use for inference, and this is nearly always sufficient to trace out a close approximation to the true marginal distribution of any subset of parameters of interest, even for missing values (which, as explained, are treated as parameters). And, if one does not have enough draws, it is simple to keep taking more, until one does.

Evaluating the Convergence of the Sampling Process

MCMC methods pose several practical questions that need to be addressed when analyzing data. What starting values should be used to initiate the sampling? How long should the cycle be (i.e., how long the burn-in period should be)? How much thinning should be done? Which algorithms will be efficient in terms of run time?

Rather than provide full answers to all these implementation issues, we will focus on one key aspect of the sampling process: the traceplot. This plot focuses on a single parameter and plots its drawn value against the iteration number. In the previously introduced case of the binomial proportion (of six divorces out of 10 couples), we use MCMC sampling to generate say 10,000 samples from a beta posterior (which arises from the conjugate beta prior and a binomial likelihood). Each of these samples represents a draw from the distribution. They can be plotted against iteration number and one can inspect whether there are systematic deviations or other obvious patterns. One looks for general stability in the traceplot, that is, for little evidence of systematic deviation (e.g., several hundred samples near $\pi = 1$, then several hundred near $\pi = 0$, both of which are endpoint values, indicating extreme deviations from a stable, interior solution). Figure 24.2 represents a well-behaved traceplot resulting from sampling from the posterior beta for our example of six divorces out of 10 couples. The sample was thinned by retaining only every 10th observation, hence 1,000 iterations are plotted out of 10,000 draws.

A more interesting example than estimating a simple proportion is using Bayesian methods to estimate the latent growth curve (we consider a real and more complex implementation of this at length at the conclusion of this chapter). This sort of model allows for two types of heterogeneity, for both slope and intercept (and higher order terms, too, given a sufficient number of time points per individual). Each subject is therefore allowed his or her own slope and intercept, but the regressions are estimated simultaneously both for efficiency and for proper modeling of the error term. The latent growth model can be estimated either in a multilevel

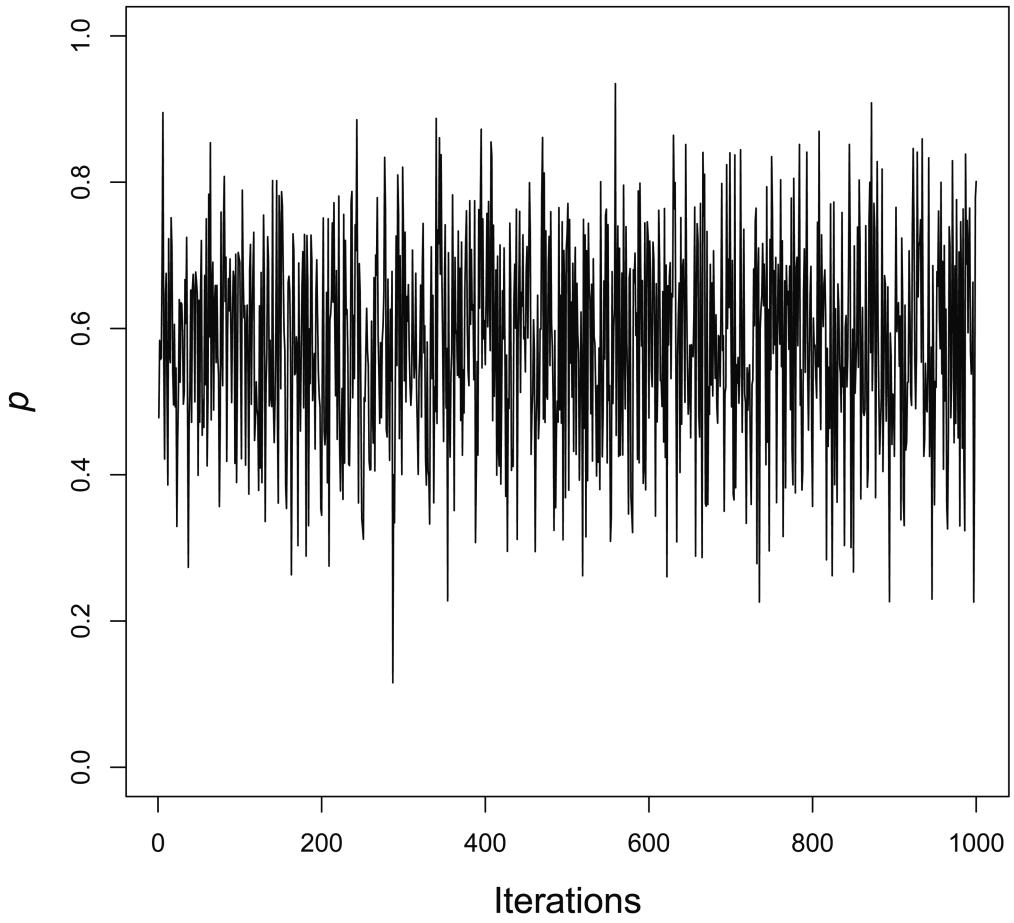

FIGURE 24.2. An example of a traceplot following the posterior density in Figure 24.1. The samples were thinned by 10. This example did not have a burn-in period because the sampling was done directly from the posterior beta with parameters 7 and 5, per the conjugate prior.

model using random terms for the slope and intercept (time points are nested within subject) or in a structural equations model using latent variables for slope and intercept (raw data are the indicators, paths are fixed to correspond to the unit vector for the intercept and the linear contrast for the slope). To illustrate, we borrowed a data set from one of our collaborators, involving four time points. We estimated a latent intercept and latent slope using the Bayesian estimation routine in Amos. The posterior distribution and the traceplot appear in Figure 24.3. We used the default in Amos of 500 iterations to burn-in, then estimated 200,000 samples, thinning by keeping every fourth, and resulting in 50,000 samples (the choice of thinning proportion is left to the researcher on the basis of the autocorrelation of the samples from the posterior; diagnostics appear in many programs to aid in this choice). The estimate of the linear latent variable is 1.286 (this is the posterior mean); the 95% (credible) interval is (1.214, 1.356). The maximum likelihood estimate for this sample is 1.285 with a standard error of 0.036 (the estimate corresponds to the fixed effect term for the slope in the multilevel model). We do not present the complete output for the other parameters, such as the intercept and the random effect variances and covariances, but similar densities and traceplots are produced for every other parameter.

Model Comparison: Bayes Factors and Deviance Information Criterion

Among the many conceptual and pragmatic difficulties of the classical approach is model comparison. In some sense, determining which model best fits given data is among the key problems in all of

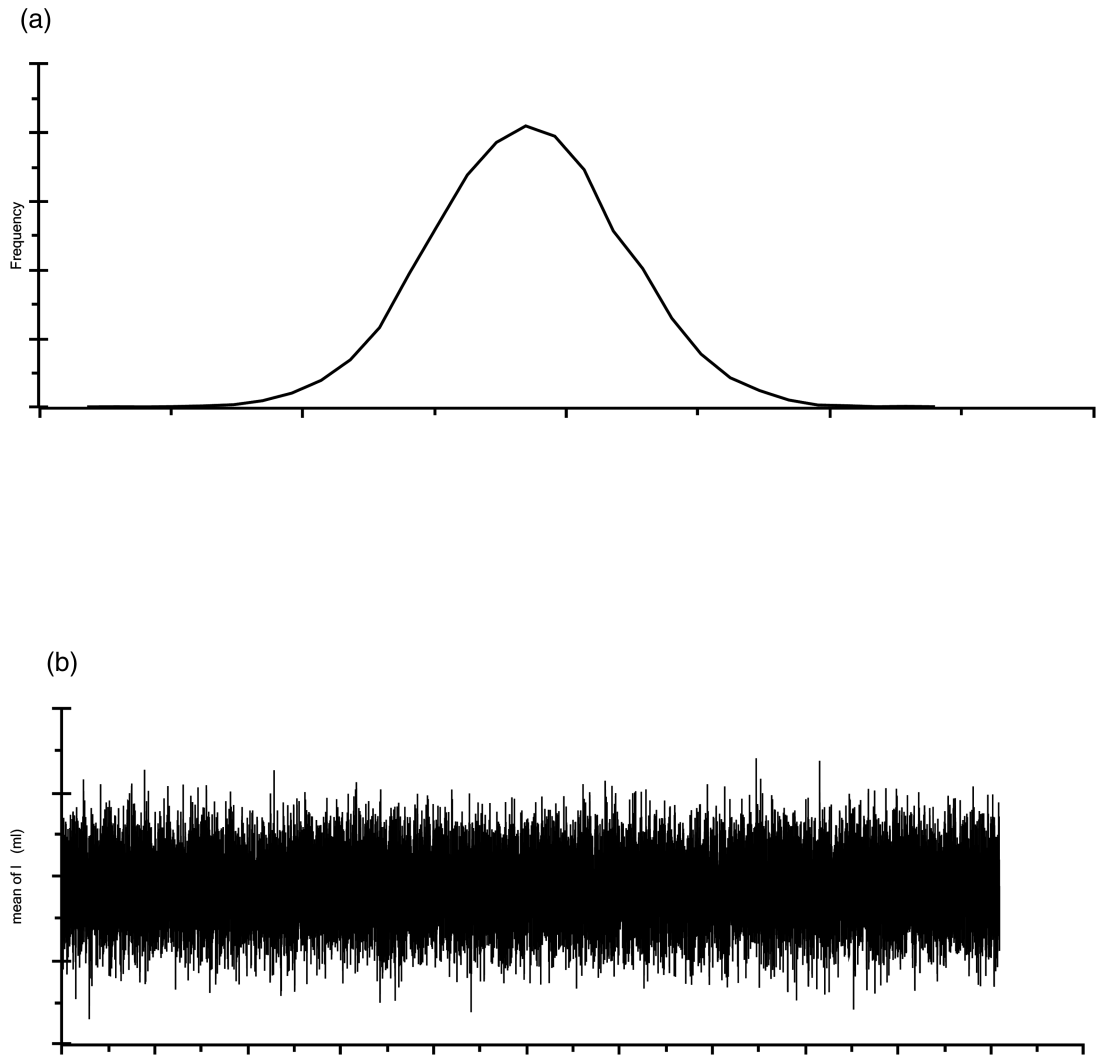

FIGURE 24.3. Example of a latent growth curve model. Results for the mean of the latent variable representing the linear term. The posterior density plot appears in panel (a); the traceplot appears in panel (b).

scientific inference. Although there are a number of specialized approaches to making this determination in classical statistics, they hold primarily for nested models. By contrast, Bayesian inference provides a general procedure for comparing *any* set of candidate models in terms of how well they are supported by the data: the Bayes factor.

Given two models, the Bayes factor quantifies how much more strongly the data favors one model over the other. Its calculation depends on one's ability to determine the so-called marginal likelihood for a model, which can be challenging in practical applications. Philosophically, the procedure is akin to the standard likelihood-ratio (LR) test in classical statistics, although the LR test depends on maximizing both model likelihoods, whereas the Bayes factor averages them over all parameters (via integrating over the prior distribution). Although it is not obvious, this one change pays great benefits: Not only can the Bayes factor compare any two candidate models but it also penalizes overparameterized models for needless complexity, unlike classical methods, which must attempt do so via various post hoc, synthetic measures such as the Akaike information

criterion and the Bayesian information criterion, none of which is overtly preferred on theoretical grounds.

Because marginal likelihoods and Bayes factors can be difficult to calculate, Bayesian statisticians have sought other comparison metrics that are easily computed from standard simulation output (i.e., the draws themselves). Among the most promising of these is the deviance information criterion (DIC). DIC sums two terms, the first assessing model lack-of-fit and the second assessing model complexity (i.e., effective number of parameters). Large values of DIC therefore indicate high lack of fit, high complexity, or both, and are therefore a less desirable model. DIC is known to be valid only when the log-likelihood is itself approximately multivariate normal, and thus must be used with caution, although it is built into many Bayesian statistical programs, simplifying model comparison considerably.

Making Predictions

It is often argued, with some justification, that real-world users of statistics have little use for parameters in and of themselves. What real users care about is using a statistical model to run what-if analyses, that is, to make predictions. Predictions can address what will happen for those units (e.g., experimental subjects, longitudinal survey respondents, and so on) already in the data or new units, the likelihood of attrition or missingness, or even the future values of parameters themselves (e.g., Are the animals becoming less sensitive to stimuli over time?). In the classical, frequentist approach, this is done via prediction intervals, at least in the standard regression or general linear models framework. But this, again, is highly dependent on asymptotic normality (or other such distributional assumptions), which may not hold for a particular data set. A secondary issue is that predictions are often made from a frequentist model using point estimates for its parameters, even though those parameters may have a complex, and relatively loose (i.e., high variance), joint distribution of their own.

Once again, the Bayesian approach supplies a complete and conceptually appealing solution to prediction: A prediction is, like everything else, simply a distribution, one that we can calculate from the posterior (not including the new observation about which we are trying to make predictions), and all available data. In simple terms, we integrate over the posterior of the parameters in the model. In symbols,

$$P\left(Y_{new} \mid Y\right) = \int P\left(Y_{new} \mid \theta\right) P\left(\theta \mid Y\right) d\theta. \qquad (4)$$

This tells us that, if we wish to know the distribution for a *new* observation, Y_{new}, we must consider all the data we already have, Y. And the way to incorporate this existing data is via the posterior probability of the parameters of the model, $P(\theta|Y)$. We simply average this (i.e., integrate) over the entire parameter space, θ. Once we have this *posterior predictive distribution*, $P(Y_{new}|Y)$, we can use it like any other distribution, to calculate means, modes, variance, quartiles, or more exotic functions. There is no guarantee that this predictive distribution will look like any of the standard distributions of elementary statistics. In fact, when this happens, it indicates that the prediction problem would have been difficult or impossible using frequentist tools alone. A simple lesson arising from this example is that the posterior distribution for the parameters, $P(\theta|Y)$, is a powerful object that can be used to readily obtain a great deal more information of use in practical statistical settings, especially so in forecasting.

Learning More About the Bayesian Approach

Many excellent textbooks provide detailed information about Bayesian inference. One of the classics is Box and Tiao (1973/1992), as much a research monograph as a textbook, which has made important contributions to several Bayesian problems. It provided much detail and explanation in deriving theoretical results in a Bayesian framework, although it did not cover modern MCMC-based approaches to Bayesian computation. Similarly, an early paper by Edwards, Lindman, and Savage (1963) made a strong case for use of Bayesian inference in psychological research.

Contemporary approaches to Bayesian estimation rely heavily on MCMC algorithms that sample the joint distribution of parameters. Such Monte Carlo techniques have been adapted to many novel model

and data types, and there are excellent textbooks on the details of various Bayesian algorithms (e.g., Robert & Casella, 2004; Tanner, 1996) as well as general introductions (e.g., Congdon, 2003; Gelman et al., 2004; Gill, 2002).

Software

Although software to implement and conduct Bayesian analyses has come about only relatively recently, many choices are presently available. We recognize that software (and associated textbook) recommendations are always a moving target, so we restrict the discussion to current capabilities, with the caveat that these will certainly deepen over time. Among the most general frameworks is the WinBUGS package (http://www.mrc-bsu.cam.ac.uk/bugs). A recent textbook teaches not only the program but also Bayesian statistics at an accessible level (Ntzoufras, 2009) and supplements two especially accessible, dedicated texts by Congdon (2003, 2007). The textbook by Gelman and Hill (2007) introduces Bayesian thinking and implementation of WinBUGS through the open-source statistics package R (http://www.r-project.org). Several SAS interfaces are available to work with WinBUGS (e.g., Smith & Richardson, 2007; Zhang, McArdle, Wang, & Hamagami, 2008), the multilevel program MLwiN has an interface to WinBUGS. A Microsoft Excel add-in, BugsXLA provides an interface to WinBUGS (http://www.axrf86.dsl.pipex.com/), and the structural equation modeling program Amos has introduced its own internal Bayesian estimation algorithm.

In addition, several books are tied to the statistical package R. These include Rossi, Allenby, and McCulloch (2005), which offers a specialized R package and several applications to marketing; the comprehensive regression textbook by Gelman and Hill (2007), which has several worked examples in R and also shows how to interface R with the dedicated Bayesian package WinBUGS; the introductory book by Gill (2002), which provides both R and WinBUGS code for standard statistical models; and the elementary book by Albert (2007), which does an exceptional job introducing theory and basic R code to implement Bayesian methods.

A welcome recent development is the inclusion of Bayesian tools in SAS, a venerable analysis platform for psychologists. By electing to include a "BAYES" statement, one can conduct Bayesian inference for a wide variety of standard specifications, most notably for generalized linear models, along with various common convergence diagnostics, like the Gelman-Rubin and Geweke. The recent addition of the MCMC procedure allows *user-specified* likelihoods and priors, with parameters that can enter the model in a linear or nonlinear functional manner. This addition literally opens the door for psychologists who wish to "go Bayesian," by allowing them to work within a software environment with which they are already comfortable.

TWO RICHER EXAMPLES ILLUSTRATING THE USEFULNESS OF THE BAYESIAN APPROACH

In this section we discuss two examples that we use to explore, at a deeper level, the concepts presented earlier in this chapter. The first is a general discussion of a canonical problem throughout the social sciences, and the second shows how a Bayesian approach can allow researchers to estimate a fairly complex model, for a real research problem, using modern-day software tools.

Multilevel Models: A Bayesian Take on a Classic Problem

We illustrate how Bayesian ideas can come into play when understanding multilevel or random effect models. Many areas of psychology have seen some form of multilevel or random effect (we will use the terms interchangeably) model come to the forefront in the past decade. Developmental psychologists use multilevel models to account for individual differences in growth-curve trajectories. Clinical psychologists use latent factors to model individual differences in scale response. Cognitive neuroscientists using functional magnetic resonance imaging in their research invoke a two-level model to account for both the intraindividual time course of the blood-oxygen-level dependence response and interindividual differences in parameters. These random effect and multilevel ideas are not new, having been

developed actively since the 1940s, if not earlier. They appear in many of the early experimental design textbooks in chapters with such titles as "Random and Nested Effects" (e.g., Winer, 1971). An important special case of this framework is the well-known repeated measures analysis of variance, in which observations are nested within subject, each subject is assigned a parameter, and data are not treated as independent observations. The correlated structure of the repeated measures is modeled through random effect terms.

Among the major limitations of the early developments in random effect and multilevel modeling was that the problem was tractable (in closed form) only for balanced designs—that is, an equal number of subjects across conditions were needed to derive formulas—and for either linear or general linear models. The major advance in the past 20 years has been the development of specialized algorithms to handle the general problem of multilevel and random effect models for a rich variety of model and data types. The new algorithms can work with unequal number of subjects (e.g., not all classrooms have to contain the same number of pupils), missing data, and so-called latent variable formulations (e.g., random utility models) and can accommodate both predictors of the random effect terms and the use of the random effect terms to predict other parameters in the model.

An important issue in working with multilevel and random effect models is that to compute estimates and standard errors, it is necessary to average over the random effect terms. That is, to estimate parameters in the classic statistical framework, it is necessary to compute the likelihood of the data at each value of the hypothesis, weight the likelihood by a function of the value of the hypothesis, and sum the products over all possible hypotheses. Typical data sets involve multiple independent observations, so the overall likelihood is taken as the product of each observation's individual likelihood. The multiplication of likelihoods (one for each observation) is justified because of the independence assumption, just as we multiply the probability of independent coin tosses to compute the joint probability of outcomes over multiple independent coin tosses. In symbols, we denote the product over multiple observations and use an integral to denote the average over the random effect term

$$\int \prod_i f\left(y_i \mid u\right) g\left(u\right) du, \tag{5}$$

where the product is taken over observations i, with likelihood $f(y_i|u)$ for a single observation, and distribution $g(u)$ over random effect u. This is a standard way to write the likelihood in the classical approach. One can then use well-known, specialized maximum likelihood techniques to estimate parameters and their standard errors directly from this likelihood (e.g., McCulloch & Searle, 2001), under suitable asymptotic assumptions.

The basic point we want to communicate is that the use of random effects involves some fairly complicated mathematical operations that do not lend themselves to easy descriptions. Expression 5 communicates the notion that there is a kind of averaging over the likelihood, where the likelihoods are weighted by the distribution $g(u)$ of the random effects. Expression 5 presents some difficult computational challenges, too. It is necessary to use specialized numerical algorithms to maximize this kind of likelihood, which contains an integral, and compute terms necessary in the classical framework, such as standard errors of the parameter estimates. There are several ways of performing a maximization over such an average, including quadrature methods and Laplace transforms, each with its pros and cons (e.g., McCulloch & Searle, 2001).

We use Expression 5 to make a simple point about the relation between Bayesian and classical methods. Expression 5 highlights a difficulty that has plagued statisticians for decades, spurring a cottage industry of ingenious computational techniques, all to more efficiently compute multilevel and random effect model parameters. Although frequentist statisticians have made great strides in surmounting the challenges that Expression 5 presents, it nonetheless entails a nasty integral, one that makes it impossible to write general, closed-form solutions, such as with unequal sample sizes or errors that are not normally distributed.

Bayesians looking at Expression 5 immediately spot a connection to a concept highly tractable within their framework. Expression 5 is proportional

to the posterior distribution (e.g., Rossi & Allenby, 2003):

$$p(u \mid y) \propto \int \prod_i f(y_i \mid u) g(u) du. \quad (6)$$

Although the classical statistician looks at the right-hand side of Equation 6 and frets about developing numerical procedures to maximize over a thorny integral, the Bayesian statistician instantly knows how to work with it, via well-established techniques for sampling from posterior distributions, such as MCMC. In addition, a set of useful tools for selecting a model, handling missing data, and assessing predictions comes along with the approach. There are a few drawbacks to the Bayesian approach. These include, for instance, having to write specialized code for specific problems (except for the simplest problems, one gives up the canned, off-the-shelf statistical package concept), work with new concepts that emerge from algorithms that use stochastic simulation, and choose a prior distribution. We do not view these as deal-breakers for using the Bayesian approach, as such issues also arise in a classical setting. For example, in a frequentist analysis, one assumes an underlying distribution and makes simplifying assumptions, such as equality of variances, to make a problem tractable; in a Bayesian setting, one selects a prior distribution. There are parallels in both cases, and in mathematical models one never completely gets away from assumptions. The key issue concerns which assumptions are more reasonable to make, which assumptions become irrelevant because of robustness issues, and which model makes difficult problems tractable.

We like this multilevel modeling example because it illustrates that there is a connection between the classical and Bayesian approaches in the case of random effect and multilevel models. The approaches turn out to be very similar: The classical statistician chooses to work with the right-hand side of Equation 6 and tackles the nasty integral directly, whereas the Bayesian chooses to work with the left-hand side, samples the posterior distribution to estimate parameters, and uses the posterior distribution to assess parametric uncertainty. They both work with the same idea; they just approach it using different methods, which we view as one of the major lessons of this chapter.

Research Example

To illustrate the power of Bayesian analysis, we present an example from recent work. We choose this example not only because it involves a data type—intent, measured on an ordinal scale—common in psychological research but also because all data and programs for analysis are freely available. The website http://cumulativetimedintent.com contains illustrative data in several formats, along with Bayesian and classical code in WinBUGS, MLwiN, and SAS, so the reader can verify directly what each approach, and program, offers in an applied context.

At the heart of the project was a need to better predict what people would purchase on the basis of their stated intentions. Studies relating intentions to behavior have been conducted for many years. The study we examine here (van Ittersum & Feinberg, 2010) introduced a new technique for eliciting individuals' intentions, by asking them to state their intent at multiple time periods on a probability scale. For example, "What is the likelihood (on an imposed 0%, 10%, 20%, . . ., 90%, 100% scale) you will have purchased this item 6 (and 12, 18, 24) months from now"? Each respondent's data looks like an increasing sequence of stated, scaled probabilities, over time. That is, can we merely *ask* people when they might purchase something and relate it, statistically, to whether and when they actually do?

In essence, this is a random effect model, but one not handled out of the box by classical estimation software. It is, however, especially amenable to Bayesian treatment. Of note for psychologists is that we can posit that each individual has some growth curve, which is taken to be linear in time (and perhaps other predictor variables as well). The key is how to relate these individual-level, latent growth curves to (a) the observable (stated probability on an ordinal scale), (b) covariates, and (c) one another. It turns out that each of these is natural in the Hierarchical Bayes approach, which is nothing more than a (nonlinear) hierarchical model, estimated using Bayesian techniques.

Suppose that the latent adoption *propensity* for subject *i* at time *t* is given by a simple linear expression,

$$Propensity_{it} = \beta_0 + \beta_{1i} t. \quad (7)$$

This specifies how the propensity changes over time for an individual but not how it varies *across* individuals. This is accomplished via a heterogeneity, or multilevel, model,

$$\beta_{1i} = \Delta z_i + u_i \quad (8)$$

$$u_i \sim N(0, \Omega_u). \quad (9)$$

This models slopes (β_{1i}) as a function of individual-level covariates (z_i) and coefficients (Δ). So-called unobserved heterogeneity, represented by u_i, is presumed normal, its degree measured by Ω_u. So far, this is exactly in keeping with standard practice in hierarchical linear models (HLM) and would in fact be equivalent to the standard formulation—which is amenable to frequentist analysis—except that we do not observe the propensity directly, but something related to it, with measurement error. Specifically, propensity, on an unbounded scale, must be functionally related to adoption probability on the unit scale. We choose a probit transform because of its conjugacy properties for Bayesian analysis:

$$\pi_{it} = \Phi(Propensity_{it}). \quad (10)$$

Because this resulting probability (π_{it}) is continuous, but our observable stated intent lies on a discrete scale, one more model stage is required. Given probability π_{it}, we can employ an especially parsimonious transformation, the *rank-ordered binomial*, to map from continuous latent, to discrete (1, . . . , *K*) observed, probabilities, which in this example has *K* = 11 (and values 0%, 10%, . . . , 100%):

$$p(Y_{it} = k) = \binom{K-1}{k-1} \pi_{it}^{k-1} (1-\pi_{it})^{K-k}, \; k = 1, \ldots, K. \quad (11)$$

Conjoining all model stages yields the following hierarchical Bayes formulation:

$$\textbf{Level I}: p(Y_{it} = k) = \binom{K-1}{k-1} \pi_{it}^{k-1} (1-\pi_{it})^{K-k}, \; k = 1, \ldots, K$$
$$Probit(\pi_{it}) = \beta_0 + \beta_{1i} t \quad (12)$$

$$\textbf{Level II}: \beta_{1i} = \Delta z_i + u_i$$
$$u_i \sim N(0, \Omega_u). \quad (13)$$

Whereas early Bayesian analyses would have required tedious specialized derivations and laborious programming, models like this one can be accommodated in dedicated software, with programs written in statistical language directly. Here, for illustration, we use MLwiN as a Bayesian computation platform (all code is posted at http://cumulativetimedintent.com). Coupled with noninformative priors, the resulting output includes samples from the posterior density for all model parameters. Automatically generated diagnostics help determine model convergence and provide plots of all marginal distributions, which do *not* have to be normal. Formal hypothesis testing proceeds off these density plots, without any distribution assumptions.

For example, we may wish to make inferences about parameters in both the Level II (heterogeneity, or dealing with the distribution of individual-level parameters) and Level I (dealing with individuals' parameters, or latent growth curves) models. Actual MLwiN output for this model, using real data, includes the following, which the program provides written in full statistical notation:

$$\text{probit}(\pi_{it}) = -1.907(0.026)\text{CONS} + \beta_{1i}\, t$$
$$\beta_{1i} = 0.733(0.056) + u_{1i}$$
$$[u_{1i}] \sim N(0, \Omega_u): \; \Omega_u = [0.508(0.068)]$$
PRIOR SPECIFICATIONS
$$P(\beta_0) \propto 1$$
$$P(\beta_1) \propto 1$$
$$p\left(\frac{1}{\sigma_{u1}^2}\right) \sim \text{Gamma}(0.001, 0.001).$$

All parts of the model are immediately recognizable as well as the estimated values of both the Level I and Level II parameters, with standard errors in parentheses. These are not merely point estimates in the usual sense, but the result of having taken 100,000 draws from the entire posterior distribution. The program automatically obtains the marginal distribution for each parameter of interest, and uses it to calculate the parameter's mean and variance, with the critical distinction that the

variance is *not* merely an approximation from the Hessian (as in frequentist analyses) but rather comes from the entire marginal distribution directly. The program also shows that it uses noninformative priors for the regression parameters (β_0 and β_1), and a mildly informative (i.e., very high variance) inverse gamma prior, a popular choice, for the variance (σ^2_{u1}).

We might interpret the model as follows. Each individual has a latent propensity to purchase (π_{it}), and the probit transform of that probability is linear in Time, with an intercept of –1.907 (*SE* = 0.026), and a coefficient (β_{1i}) with a mean of 0.733 (*SE* = 0.056). However, there is some degree of *variation* in the value of this coefficient across respondents. The mean across respondents, as we have seen, is 0.733, but the variance is estimated to be 0.508 (standard error: 0.068). We would also wish to check that the traceplot for each of these parameters looked reasonable, meaning like a sequence of independent draws, with no patterns obvious to the eye. These appear in Figure 24.4, also as generated automatically in MLwiN, for the last 10,000 draws for each parameter (we have also included a kernel density for the variance and, as would be expected, it is not symmetric). Because we can access all these

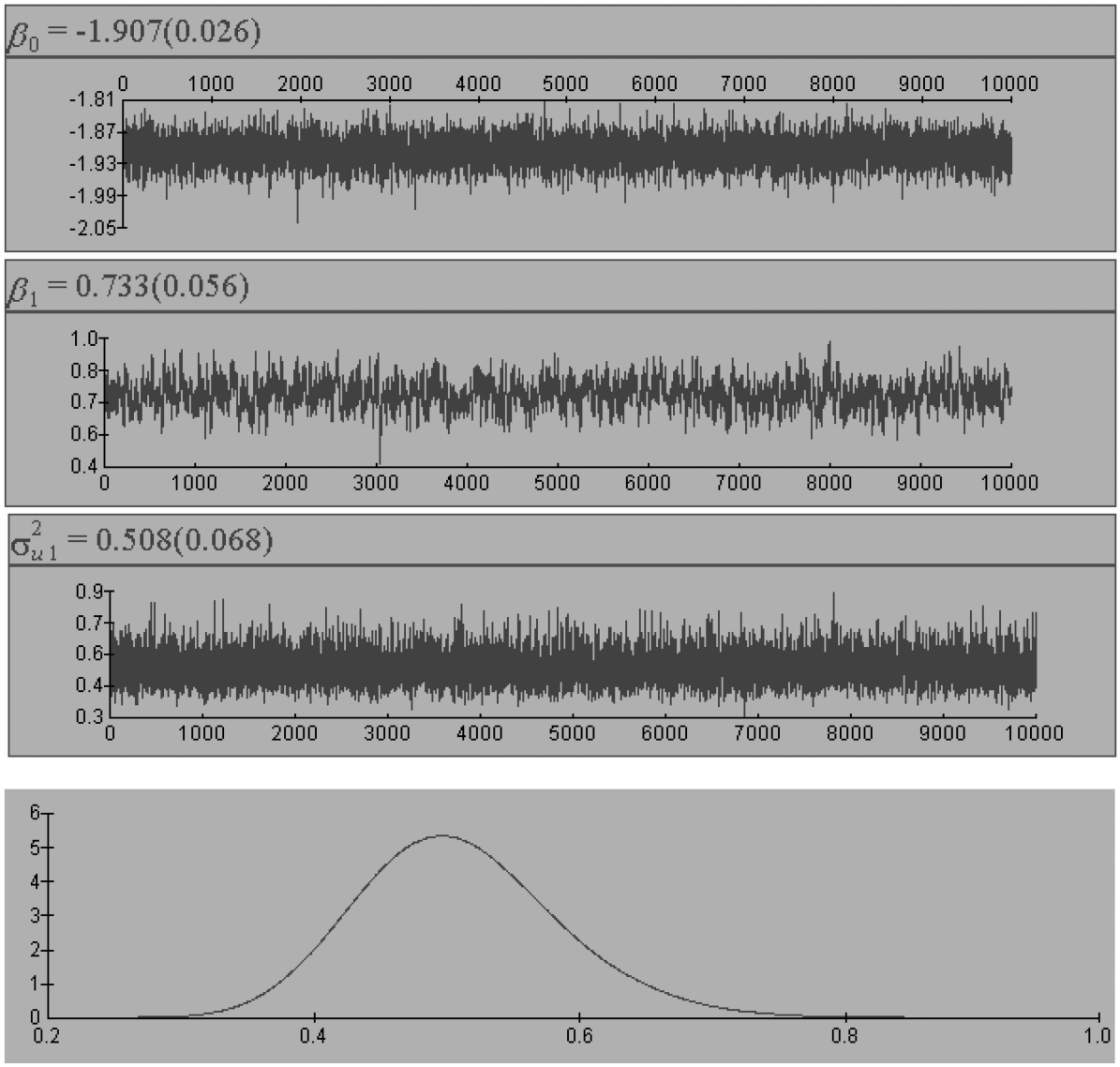

FIGURE 24.4. Traceplots for three parameters from a latent purchase intent model, with a kernel density for the variance, generated in MLwiN.

draws, and indeed have them for each of the respondent's individual slope coefficients (β_{1i}), we can calculate latent growth curves for each, error bars around them, and in fact *any* function of their parameters, all by operating on the posterior draws directly.

Although this model would not be impossible to estimate using classical techniques—indeed, one can program it using PROC NLMIXED in SAS, with some patience, by writing out the model likelihood directly—the Bayesian approach allows all parameters of interest to be calculated to any desired degree of accuracy. Moreover, we obtain a *distribution* for each of these parameters, and an arbitrarily large number of draws from each one. In practical terms, this means that the analyst is freed from making any assumptions about the asymptotic behavior of parameters and can perform on-the-fly postestimation tests on complex functions of the problem's parameters. This is completely beyond what frequentist techniques can offer, yet it is natural and straightforward using Bayesian estimation.

CONCLUSION

Some readers will get the sense that our views about Bayesian statistics are not entirely mainstream. Partisans will undoubtedly feel we did not portray their vantage point with sufficient detail. The classical statistician may take issue with superficial attention to the problem of defining one's prior. "Ambiguity over selecting a prior distribution is the Achilles' heel of the Bayesian approach," a classically inclined researcher may say. Bayesians may be incensed that we lump their elegant, comprehensive formalism with the classical approach by saying they both act the same when the Bayesian assumes a noninformative prior. "But you miss the important differences between how we interpret the results," will be shouted from the Bayesian rooftops. Let us be the first to acknowledge that some of the subtle details have been omitted. But that was completely intentional. We want to bring more researchers to the discussion, expose more people to the underpinnings of both classical and Bayesian approaches, and show researchers some new tools. We believe (and we have a pretty sharp prior on that belief) that the best way to accomplish this is by outlining the similarities of the approaches and the advantages each offers.

We hope this chapter has been a readable and accessible introduction to the basic notions of Bayesian statistics and that it provides a straightforward way to formulate some of the tools that the Bayesian tradition offers. In these relatively few pages we cannot cover all the ins and outs of conducting different types of Bayesian analyses—there are books that do that. If the reader's interest is piqued sufficiently to seek out some of the reference books and explore some of the software we mention, then this chapter has been successful.

Some areas of psychology have already started to apply modern Bayesian methods. For example, new models in item response theory have used Bayesian ideas to estimate multivariate, multilevel, second-order, item-response theory models (e.g., Duncan & MacEachern, 2008; Fox & Glas, 2001; Sheng & Wikle, 2008). We hope these and other examples will provide the inspiration to seek new ways to test your research ideas and that Bayesian methods provide some useful tools to carry out those tests.

References

Albert, J. (2007). *Bayesian computation with R*. New York, NY: Springer. doi:10.1007/978-0-387-71385-4

Box, G. E. P., & Tiao, G. C. (1992). *Bayesian inference in statistical analysis*. New York, NY: Wiley. (Original work published 1973)

Congdon, P. (2003). *Applied Bayesian modelling*. Chichester, England: Wiley. doi:10.1002/0470867159

Congdon, P. (2007). *Bayesian statistical modelling* (2nd ed.). Chichester, England: Wiley.

Duncan, K., & MacEachern, S. (2008). Nonparametric Bayesian modeling for item response. *Statistical Modelling*, *8*, 41–66. doi:10.1177/1471082X0700800104

Edwards, W., Lindman, H., & Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242. doi:10.1037/h0044139

Fox, J.-P., & Glas, C. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271–288. doi:10.1007/BF02294839

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. Boca Raton, FL: Chapman & Hall.

Jaynes, E. (1986). Bayesian methods: General background. In J. H. Justice (Ed.), *Maximum entropy and Bayesian methods in applied statistics* (pp. 1–25). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511569678.003

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York, NY: Wiley.

Ntzoufras, I. (2009). *Bayesian modeling using WINBUGS*. Hoboken, NJ: Wiley. doi:10.1002/9780470434567

Robert, C., & Casella, G. (2004). *Monte Carlo statistical methods*. New York, NY: Springer.

Rossi, P. E., & Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science*, *22*, 304–328. doi:10.1287/mksc.22.3.304.17739

Rossi, P. E., Allenby, G. M., & McCulloch, R. (2005). *Bayesian statistics and marketing*. Chichester, England: Wiley.

Sheng, Y., & Wikle, C. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, *68*, 413–430. doi:10.1177/0013164407308512

Smith, M. K., & Richardson, H. (2007). WinBUGSio: A SAS macro for the remote execution of WinBUGS. *Journal of Statistical Software*, *23*, 1–10.

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions*. New York, NY: Springer.

van Ittersum, K., & Feinberg, F. M. (2010). Cumulative timed intent: A new predictive tool for technology adoption. *Journal of Marketing Research, 47*, 808–822.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York, NY: McGraw-Hill.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*, 301–322. doi:10.1037/a0016972

Zhang, Z., McArdle, J., Wang, L., & Hamagami, F. (2008). A SAS interface for Bayesian analysis with WinBUGS. *Structural Equation Modeling*, *15*, 705–728. doi:10.1080/10705510802339106