16. Schoenfeld, D. (1982) Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239–241.
17. Lin, D. Y. Wei, L. J., and Ying, Z. (1993) Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*. **80**, 557–572.
18. Therneau, T. M., and Grambsch, P. M. (2000) Modeling survival data: extending the Cox model. Berlin, New York, Springer-Verlag.
19. Freireich, E. J., Gehan, E., Frei, E., Schroeder, L. R., Wolman, I. J., Anbari, R., Burgert, E. O., Mills, S. D., Pinkel, D., Selawry, O. S., and others. (1963) The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: a model for evaluation of other potentially useful therapy. *Blood* **21**, 699–716.

16

Basic Bayesian Methods

Mark E. Glickman and David A. van Dyk

Summary

In this chapter, we introduce the basics of Bayesian data analysis. The key ingredients to a Bayesian analysis are the likelihood function, which reflects information about the parameters contained in the data, and the prior distribution, which quantifies what is known about the parameters before observing data. The prior distribution and likelihood can be easily combined to from the posterior distribution, which represents total knowledge about the parameters after the data have been observed. Simple summaries of this distribution can be used to isolate quantities of interest and ultimately to draw substantive conclusions. We illustrate each of these steps of a typical Bayesian analysis using three biomedical examples and briefly discuss more advanced topics, including prediction, Monte Carlo computational methods, and multilevel models.

Key Words: Monte Carlo simulation; posterior distribution; prior distribution; subjective probability.

1. Introduction

As with most academic disciplines, researchers and practitioners often choose from among several competing schools of thought. In music, for example, some composers have been guided by the rules of Romanticism, Impressionism, or Atonality in developing their work; in art, painters have at various periods followed the rules of Cubism, Expressionism, or Dadaism with widely differing results. One might assume that a scientific discipline such as statistics is immune to such philosophical divides. Interestingly, this is not the case. Statistics, as a discipline, consists of two main competing schools of thought: The *frequentist* or classical approach to statistical inference, and the *Bayesian* approach. The frequentist approach, which includes hypothesis testing and confidence intervals as two of the main modes of inference, has been the main framework for

From: *Methods in Molecular Biology, vol. 404: Topics in Biostatistics*
Edited by: W. T. Ambrosius © Humana Press Inc., Totowa, NJ

most of the techniques discussed thus far in this book. We discuss the basics of the Bayesian approach in this chapter.

The underlying difference between the Bayesian and frequentist approaches to statistical inference is in the definition of probability. A frequentist views probability as a long-run frequency. When a frequentist asserts that the probability of a fair coin tossed landing heads is $\frac{1}{2}$, he means that in the long run, over repeated tosses, the coin will land heads half the time. In contrast, a Bayesian, who will also surely say that the probability a coin lands heads is $\frac{1}{2}$, is expressing a *degree of belief* that the coin lands heads, perhaps arguing that based on the symmetry of the coin there is no reason to think that one side is more likely to come up than the other side. This definition of probability is usually termed *subjective* probability. Whereas, in practice, a frequentist uses probability to express the frequency of certain types of data to occur over repeated trials, a Bayesian uses probability to express belief in a statement about unknown quantities.

These definitions have profound impact on a framework for statistical inference. Because a Bayesian uses subjective probability, he can describe uncertainty of a statement about an unknown parameter in terms of probability. A frequentist cannot. So, for example, it is legitimate for a Bayesian to conclude as a result of a data analysis that an interval contains a parameter of interest with 95% probability. A frequentist, in contrast, will use probability to describe how often the calculations that produce an interval will cover the parameter of interest in repeated samples. For instance, frequentist 95% confidence intervals have the property that, in the long run, 95% of such intervals will cover the parameters being estimated. But, unfortunately for the frequentist, once a set of data is observed and an interval is computed, the frequentist concept of probability is no longer relevant. Further, when a Bayesian is evaluating two competing hypotheses about an unknown parameter, he can calculate the probability of each hypothesis given observed data and then choose the hypothesis with the greater probability. A frequentist, on the other hand, cannot use probability in such a direct way, and instead will approach the problem asymmetrically and ponder the long-run frequency under one of the hypotheses of sampling data as extreme or more extreme than what was observed.

This chapter describes the basics of Bayesian statistics. We begin by describing the main ingredients of a Bayesian analysis. In this discussion, we explain how to obtain the *posterior distribution* of model parameters and how to obtain useful model summaries and predictions for future data. We then demonstrate an application of the Bayesian approach to multilevel models, using *Monte Carlo simulation* as a computational tool to obtain model summaries.

2. Fundamentals of a Bayesian Analysis

A typical Bayesian analysis can be outlined in the following steps.

1. Formulate a probability model for the data.
2. Decide on a *prior distribution*, which quantifies the uncertainty in the values of the unknown model parameters *before* the data are observed.
3. Observe the data, and construct the *likelihood* function (see **Section 2.3**) based on the data and the probability model formulated in **step 1**. The likelihood is then combined with the prior distribution from **step 2** to determine the posterior distribution, which quantifies the uncertainty in the values of the unknown model parameters *after* the data are observed.
4. Summarize important features of the posterior distribution, or calculate quantities of interest based on the posterior distribution. These quantities constitute statistical outputs, such as point estimates and intervals.

We discuss each of these steps in turn in **Sections 2.1–2.4**.

The main goal of a typical Bayesian statistical analysis is to obtain the posterior distribution of model parameters. The posterior distribution can best be understood as a weighted average between knowledge about the parameters before data is observed (which is represented by the prior distribution) and the information about the parameters contained in the observed data (which is represented by the likelihood function). From a Bayesian perspective, just about any inferential question can be answered through an appropriate analysis of the posterior distribution. Once the posterior distribution has been obtained, one can compute point and interval estimates of parameters, prediction inference for future data, and probabilistic evaluation of hypotheses. Predictive inference is the topic of **Section 2.5**.

2.1. Data Models

The first step in a Bayesian analysis is to choose a probability model for the data. This process, which is analogous to the classic approach of choosing a data model, involves deciding on a probability distribution for the data if the parameters were known. If the n data values to be observed are $y_1, \ldots, y_n$, and the vector of unknown parameters is denoted $\boldsymbol{\theta}$, then, assuming that the observations are made independently, we are interested in choosing a probability function $p(y_i \mid \boldsymbol{\theta})$ for the data (the vertical bar means "conditional on" the quantities to the right). In situations where we have extra covariate information, x_i, for the ith case, as in regression models, we would choose a probability function of the form $p(y_i \mid x_i, \boldsymbol{\theta})$. When the data are not conditionally independent given the parameters and covariates, we must specify the joint probability function, $p(y_1, \ldots, y_n \mid x_1, \ldots, x_n, \boldsymbol{\theta})$.

Example 1

A random sample of 300 women aged 60–69 years whose immediate families have had histories of cancer are to be screened for breast cancer. Let y_i be 1 if woman i has a positive test, and 0 if not, for $i = 1, \ldots, 300$. Let θ be the probability that a randomly selected woman aged 60–69 years with a family history of cancer has a positive breast cancer screening. Then an appropriate model for the data is to assume that the y_i independently follow a Bernoulli distribution with probability θ, that is,

$$p(y_i \mid \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

for $i = 1, \ldots, 300$.

Example 2

A random sample of 50 men with a history of cardiovascular disease enters a study on LDL (low-density lipoprotein) cholesterol. Let y_i be the LDL cholesterol level (in mg/dL) for man i, $i = 1, \ldots, 50$. A reasonable probability model for LDL cholesterol levels is a normal distribution. We can assume that the y_i are independently normal with unknown common mean μ and variance σ^2. The probability function for y_i is given by

$$p(y_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(y_i-\mu)^2/2\sigma^2\right)$$

for $i = 1, \ldots, 50$.

2.2. Prior Distribution

Once the data model (probability model) is chosen, a Bayesian analysis requires the assertion of a prior distribution for the unknown model parameters. The prior distribution can be viewed as representing the current state of knowledge, or current description of uncertainty, about the model parameters prior to data being observed.

Approaches to choosing a prior distribution divide into two main categories. The first approach involves choosing an *informative* prior distribution. With this strategy, the statistician uses his knowledge about the substantive problem perhaps based on other data, along with elicited expert opinion if possible, to construct a prior distribution that properly reflects his (and experts') beliefs about the unknown parameters. The notion of an informative prior distribution may seem at first to be overly subjective and unscientific. In response to this concern, it should be pointed out that the selection of a data model, which a frequentist needs to make, is also a subjective choice, so that frequentist analyses are not devoid from subjectivity either. Furthermore, it can be argued that if extra information or knowledge about the model parameters exists prior to observing data, it would be unscientific *not* to incorporate such information into a data analysis. For example, in a study measuring the weight of preterm births, it would be sensible to incorporate into the prior distribution that the "prior probability" of a mean birth weight above 15 lb is negligible. Another criticism by

frequentists of using informative prior distributions is that two Bayesian statisticians are likely to use two different prior distributions, which leads to two different sets of inferences for the same scientific problem. Again, it is reasonable to respond to this criticism by pointing out that when frequentists use different data models on the same data, conclusions will be different as well. From a Bayesian point of view, a prior distribution is part of the overall statistical model, so that two Bayesian statisticians selecting different prior distributions is analogous to two frequentist statisticians choosing two different data models.

The second main approach to choosing a prior distribution is to construct a *noninformative* prior distribution that represents ignorance about the model parameters. Besides *noninformative*, this type of distribution is also called objective, vague and diffuse, and sometimes a reference prior distribution. Choosing a noninformative prior distribution is an attempt at objectivity by acting as though no prior knowledge about the parameters exists before observing the data. This is implemented by assigning equal probability to all values of the parameter (or at least approximately equal probability over localized ranges of the parameter). The appeal of this approach is that it directly addresses the criticisms of informative prior distributions as being subjectively chosen. In some cases, there is arguably a single best noninformative prior distribution for a given data model, so that this prior distribution can be used as a default option, much like one might have default arguments in computer programs. Unfortunately, noninformative prior distributions are not without their problems either. First, because there are various commonly accepted criteria for constructing noninformative prior distributions, it is rare that, for a given data model, all these criteria produce the same unique noninformative prior distribution. Second, some common methods for constructing noninformative prior distributions, such as always assuming a uniform distribution for a parameter, result in an interesting inconsistency. Any method for constructing a noninformative prior distribution ought to be invariant to the measurement scale of the parameter; if, for example, the method of constructing a noninformative prior distribution is applied to a data model with parameter θ, and then applied to the same model reparameterized with parameter $\eta = \log(\theta)$, it would be desirable that the distributions on θ and η were representing equivalent probabilistic information. It turns out that this is a difficult criterion to satisfy (one approach constructed to satisfy this invariance criterion is *Jeffrey's rule*, which works well with one-parameter data models but with mixed results for multiparameter models). Finally, many commonly used methods for constructing a noninformative prior distribution result in probability functions that integrate to infinity, usually called *improper* distributions, and are not formally probability distributions. Luckily, for many problems, having an improper prior distribution still allows for a coherent Bayesian analysis.

In general, if an objective prior distribution is desired, one defensible strategy is to construct a relatively uniform proper (i.e., integrates to 1) prior distribution. If the information contained in the data is supposed to be the main determining factor in producing statistical inferences (as it should be), then we should expect that the choice among a range of relatively flat prior distributions will not make much of a difference. On the other hand, if the choice of a relatively flat prior distribution does matter, this may be an indication that the data conveys little information about the parameter of interest, and it may be appropriate to rethink the form of the data model, or to collect additional data.

Example 1 (Continued)

Recall that θ is the probability a randomly selected woman, aged 60–69 years with a family history of cancer, has a positive breast cancer screening. According to the American Cancer Society, roughly 3.6% of women aged 60–69 years develop invasive breast cancer, so that we may form an informative prior distribution for θ that reflects this information. A flexible choice of a prior distribution for a Bernoulli probability is $\theta \sim \text{Beta}(\alpha, \beta)$, that is, θ has a Beta distribution with specified parameters α and β. The probability function is given by

$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where $\Gamma()$ represents the Gamma function.[1] The mean of a Beta distribution is $\alpha/(\alpha+\beta)$. The value $\alpha+\beta$ has an interpretation as the amount of information about θ viewed as a sample size. For the cancer screening problem, the choice $\theta \sim \text{Beta}(0.36, 9.64)$ is sensible, as this distribution has a mean of $0.36/(0.36 + 9.64) = 0.036$, the estimate given by the American Cancer Society, and the information represented by this distribution is equivalent to that in $0.36 + 9.64 = 10$ data values. A plot of the probability function is given in **Figure 1**. Note that the greatest probability under this distribution of θ is concentrated around very low values, which is meant to reflect our initial belief that a value of θ much larger than 0.1 or 0.15 is not very plausible. With an eventual sample of 500 observations, the data is about 50 times more informative than the prior distribution.

Example 2 (Continued)

For studying LDL cholesterol levels, we assume a noninformative prior distribution for the mean μ and variance σ^2 of the normal data model. A strategy that can often be employed for models with multiple parameters is to consider each parameter separately and form the joint prior distribution as a product of the several independent distributions.

The most common noninformative choice for a location parameter, such as a mean (or a regression coefficient), is to assume an *improper* uniform distribution over the entire real line. Thus we assume

$$p(\mu) = 1$$

for $-\infty < \mu < \infty$ even though this function does not integrate over the range. We further assume, independently, that the prior distribution for σ^2 is the improper probability function

$$p(\sigma^2) = 1/\sigma^2.$$

By a change-of-variables argument from elementary calculus, this distribution on σ^2 corresponds with a uniform distribution on $\log(\sigma^2)$ over the entire real line. Besides having the appeal of placing a uniform distribution over a parameter that has been transformed to take values over the entire real line, as with μ, this prior distribution also recognizes that extremely large values of σ^2 are less believable *a priori* than are small values. A uniform distribution on the untransformed variance, σ^2, in contrast, asserts that a variance between 1,000,000 and 1,000,001 is as likely *a priori* as a variance between zero and one, which is not particularly believable. We therefore assume an improper joint prior distribution for (μ, σ^2) equal to

$$p(\mu, \sigma^2) = p(\mu)p(\sigma^2) = 1 \cdot (1/\sigma^2) = 1/\sigma^2.$$

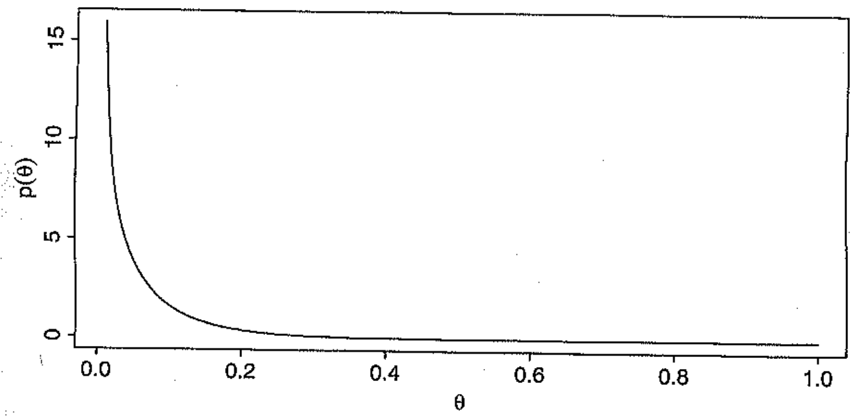

Fig. 1. Probability function for the Beta(0.36, 9.64) distribution.

[1] The Gamma function is closely related to the factorial function: For a positive integer n, $\Gamma(n) = (n-1)!$. For more details about the Gamma function, see *(1)*.

2.3. From the Likelihood to the Posterior Distribution

Once the data has been observed, the likelihood function, or simply the likelihood, is constructed. The likelihood is the joint probability function of the data, but viewed as a function of the parameters, treating the observed data as fixed quantities. Assuming that the data values, $\mathbf{y} = (y_1, \ldots, y_n)$ are obtained independently, the likelihood function is given by

$$L(\theta | y) = p(y_1, \ldots, y_n | \theta) = \prod_{i=1}^{n} p(y_i | \theta).$$

In the Bayesian framework, all of the information about $\boldsymbol{\theta}$ coming directly from the data is contained in the likelihood. Values of the parameters that correspond with the largest values of the likelihood are the parameters that are most supported by the data.

To obtain the posterior distribution, $p(\boldsymbol{\theta}|\mathbf{y})$, the probability distribution of the parameters once the data have been observed, we apply Bayes' theorem:

$$p(\boldsymbol{\theta}|y) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{\theta}|y)}{\int p(\boldsymbol{\theta})p(y|\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|y)}{p(y)} \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|y)$$

where "$\propto$" means "is proportional to" (i.e., that the expressions are equal when the right-most term is multiplied by a normalizing constant that doesn't depend on $\boldsymbol{\theta}$). Operationally, therefore, it is straightforward in principle to obtain the posterior distribution: Simply multiply the prior distribution by the likelihood, and then determine the constant (not depending on $\boldsymbol{\theta}$) that forces the expression to integrate to 1. An effective strategy for computing the posterior distribution is to drop multiplicative constants from the prior distribution and likelihood that do not depend on $\boldsymbol{\theta}$, and then in the final step determine the normalizing constant.

Example 1 (Continued)

Suppose, for the breast cancer screening study, 14 of the 300 women had positive tests. Thus 14 women have $y_i = 1$, and the remaining 286 have $y_i = 0$. The likelihood is therefore given by

$$L(\theta | y) = \prod_{i=1}^{300} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{14}(1-\theta)^{286}.$$

The posterior distribution is proportional to the product of the Beta prior distribution (with parameters $\alpha = 0.36$ and $\beta = 9.64$) and the likelihood,

$$L(\theta|y) \propto p(\theta)L(\theta|y) \propto \left(\frac{\Gamma(10)}{\Gamma(0.36)\Gamma(9.64)}\theta^{-0.64}(1-\theta)^{8.64}\right)\cdot\theta^{14}(1-\theta)^{286}$$
$$\propto \theta^{-0.64}(1-\theta)^{8.64}\cdot\theta^{14}(1-\theta)^{286} \propto \theta^{13.36}(1-\theta)^{294.64}.$$

Note that the normalizing constant in the prior distribution was dropped as it does not depend on θ. Rather than determine the normalizing constant analytically, we notice that the final expression is proportional to a Beta distribution with parameters $\alpha = 14.36$ and $\beta = 295.64$, so that the posterior distribution must be

$$p(\theta | y) = \frac{\Gamma(330)}{\Gamma(14.36)\Gamma(295.64)}\theta^{13.36}(1-\theta)^{294.64}.$$

Thus, the posterior distribution is $\theta|\mathbf{y} \sim \text{Beta}(14.36, 295.64)$.

Example 2 (Continued)

In the LDL cholesterol study, suppose the 50 LDL cholesterol measurements are taken. The likelihood is the product of 50 normal probability functions:

$$L(\mu, \sigma^2 | y) = \prod_{i=1}^{50} p(y_i | \mu, \sigma^2) = \prod_{i=1}^{50} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(y_i - \mu)^2/2\sigma^2)$$
$$= \frac{1}{(2\pi\sigma^2)^{25}} \exp\left(-\sum_{i=1}^{50}(y_i - \mu)^2/2\sigma^2\right).$$

Letting $\bar{y} = \frac{1}{n}\sum_{i=1}^{50} y_i$ and $s^2 = \frac{1}{49}\sum_{i=1}^{50}(y_i - \bar{y})^2$ be the sample mean and variance, respectively, the likelihood can be rewritten in a more useful form as

$$L(\mu, \sigma^2 | y) = \frac{1}{(2\pi\sigma^2)^{25}} \exp\left(-\sum_{i=1}^{50}(y_i - \mu)^2/2\sigma^2\right)$$
$$= \frac{1}{(2\pi\sigma^2)^{25}} \exp(49s^2 + 50(\mu - \bar{y})^2/2\sigma^2).$$

We again use the standard choice of noninformative prior distribution on the parameters of a normal model, $p(\mu, \sigma^2) = 1/\sigma^2$. With this choice of prior distribution, the posterior distribution can be computed as follows:

$$p(\mu, \sigma^2 | y) \propto p(\mu, \sigma^2)L(\mu, \sigma^2 | y) \propto \frac{1}{\sigma^2}\cdot\frac{1}{(2\pi\sigma^2)^{25}} \exp(-(49s^2 + 50(\mu - \bar{y})^2)/2\sigma^2)$$
$$\propto (\sigma^2)^{-25.5}\exp(-49s^2/2\sigma^2)\cdot\frac{1}{\sigma}\exp\left(-(\mu - \bar{y})^2/2(\sigma/\sqrt{50})^2\right).$$

The second term in the above expression, as a function of μ with the appropriate constant, is a normal distribution with mean $\bar{y}$ and variance $\sigma^2/50$. The first term, with the appropriate constant, is an *inverse-*χ^2 distribution; this means that $1/\sigma^2$ has the more familiar chi-square distribution. The posterior distribution $p(\mu, \sigma^2|y)$ therefore factors into a *marginal posterior* distribution of σ^2, $p(\sigma^2|y)$, which is inverse-χ^2, and a *conditional posterior* distribution of μ given σ^2, $p(\mu|\sigma^2, y)$, which is normal. A marginal posterior distribution specifies the posterior distribution for a subset of the model parameters without regard to the other parameters. A conditional posterior distribution, on the other hand, is the posterior distribution of a subset of the parameters subject to the other parameters having specified values.

In this example, the joint posterior distribution can be written

$$p(\mu, \sigma^2|y) = p(\sigma^2|y)p(\mu|\sigma^2, y)$$

where $\sigma^2|\mathbf{y} \sim \text{Inv-}\chi^2(49, s^2)$ (i.e., $49s^2/\sigma^2$ has a chi-square distribution on 49 degrees of freedom), and $\mu|\sigma^2, \mathbf{y} \sim \text{N}(\bar{y}, \sigma^2/50)$. Once the sample mean and sample variance have been computed from the data, these values can be substituted in to obtain the actual distributions. It is also worth noting that σ^2 can be integrated out of the joint posterior density to obtain the *marginal* posterior density of μ, which is

$$\mu|y \sim t_{49}(\bar{y}, s^2/50),$$

that is, a t-distribution with 49 degrees of freedom that is centered at $\bar{y}$ and rescaled by $s/\sqrt{50}$.

2.4. Posterior Summaries

Once the posterior distribution has been determined, inferential conclusions can be summarized with an appropriate analysis. Point estimates of parameters are commonly computed as the mean or the *mode* (i.e., highest point) of the posterior distribution. Interval estimates can be calculated by producing the end points of an interval that correspond with specified percentiles of the posterior distribution. For example, a 95% *central posterior interval* involves computing the 2.5%-ile and 97.5%-ile of the posterior distribution. Probabilities of competing composite hypotheses can be evaluated by calculating their posterior probability, that is, the probability of the hypotheses based on the posterior distribution.

Example 1 (Continued)

With a posterior distribution for the probability of a positive breast cancer screening of Beta(14.36, 295.64), we can compute informative inferential summaries about θ. The posterior mean and posterior mode are the two most common point estimates for a parameter. For a Beta distribution with parameters α and β, the mean is $\alpha/(\alpha+\beta)$, and the mode is $(\alpha-1)/(\alpha+\beta-2)$. The posterior mean estimate of θ is therefore

$$\text{E}(\theta|\mathbf{y}) = 14.36/(14.36 + 295.64) = 0.0463.$$

The posterior mode estimate of θ, the most "believable" value of θ, is

$$\text{Mode}(\theta|\mathbf{y}) = (14.36 - 1)/(14.36 + 295.64 - 2) = 0.0434.$$

To construct a 95% central posterior interval for θ, we need to find the appropriate percentiles of the Beta(14.36, 295.64) distribution. Analytically, this involves evaluating the integral $\int_0^c p(\theta|\mathbf{y})d\theta = 0.025$ and solving for c to obtain the lower end point of the interval, and similarly for the upper end point. Using statistical software (like R or S-Plus, SAS, Stata, SPSS, etc.), the percentiles can easily be evaluated numerically. The 2.5%-ile and the 97.5%-ile of the posterior distribution are computed to be 0.0259 and 0.0723, respectively, so that the 95% central posterior interval for θ is (0.0259, 0.0723). There is a 0.95 posterior probability that θ lies in this interval.

Suppose for health policy reasons that it is important to know whether $\theta > 0.05$. We can translate the question into a posterior probability computation of

$$\text{P}(\theta > 0.05|y) = \int_{0.05}^{1} p(\theta|y)\,d\theta.$$

Rather than attempting to evaluate this Beta integral analytically, we can evaluate it numerically using statistical software. The probability from the Beta posterior distribution is computed to be 0.351, which implies that the probability $\theta < 0.05$ is 0.649. Thus we may conclude that it is more likely than not that $\theta < 0.05$.

Example 2 (Continued)

We computed the joint posterior distribution of μ and σ^2, the mean and variance of the normal model, in the LDL cholesterol study. This posterior distribution depends on the data through the sample mean and sample variance of the 50 measurements, $\bar{y}$ and s^2, respectively. Now suppose that upon observing the measurements, we compute $\bar{y} = 110$ and $s^2 = 100$. From a Bayesian perspective, the posterior distribution is a complete summary of what we know about the parameters, both from the data and—as quantified via the prior distribution—from other sources of information. In this case, we can plot the posterior distribution and use the plots to quantify what we understand about the unknown

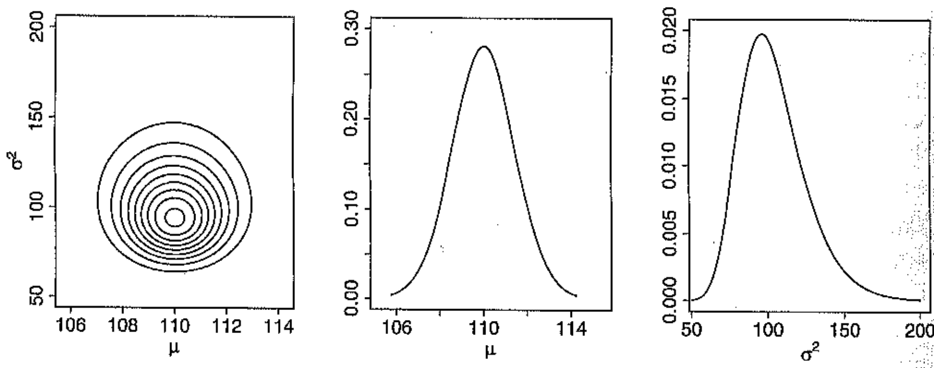

Fig. 2. The posterior distribution of parameters of LDL cholesterol levels. The three figures depict the 2-dimensional joint posterior distribution of the mean and variance of LDL cholesterol in the population of men. A contour plot of the joint distribution and plots of both of the marginal distributions are given.

parameters. A contour plot of the joint posterior distribution appears in the first panel of **Figure 2.** The next two panels represent the marginal posterior distributions of μ and σ^2, respectively. These distributions represent our knowledge about likely values of the mean and variance of LDL cholesterol levels in this particular population of men. Judging from the posterior distribution of μ, the mean LDL cholesterol level is about 110 plus or minus about four. The posterior distribution of σ^2 tells us how much the level varies among men: The variance appears to be about 100 but could be as low as 60 or as high as 175. Notice that the posterior distribution of σ^2 is slightly skewed toward the right. Looking at the joint distribution, the mean and variance appear to be uncorrelated. This means that inference about particular values of μ does not have a relationship to our inference about values of σ^2.

2.5. Predictive Distributions

One of the benefits of the Bayesian approach is that predictive inference is a straightforward computation once the posterior distribution has been obtained. Suppose we have observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$, and we would like to make a prediction about a future observation y. From an analysis of the data, we have obtained $p(\boldsymbol{\theta}|\boldsymbol{y})$, the posterior distribution. We are interested in making probabilistic statements about an unobserved y, so that we want to compute the *posterior predictive distribution* of y. The posterior predictive distribution is written as $p(y \mid \boldsymbol{y})$. Note that we are not interested in conditioning on parameter values, but that we only want to condition on what we have observed: the previous data.

The posterior predictive distribution can be computed using the equation

$$p(y|\boldsymbol{y})=\int p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$

which makes the often appropriate assumption that future data is independent of past data conditional on the parameters. Thus, integrating the product of the data model distribution with the posterior distribution with respect to the model parameters produces the posterior predictive distribution, which can then be summarized for predictive inferences.

Example 2 (Continued)

Let y be an LDL cholesterol measurement taken of a man with a history of cardiovascular disease not yet sampled. We are interested in deriving the posterior predictive distribution of y, that is, $p(y|\boldsymbol{y})$. We must therefore evaluate

$$\begin{aligned} p(y|\boldsymbol{y}) &= \iint p(y|\mu,\sigma^2)p(\mu,\sigma^2|\boldsymbol{y})\,d\mu d\sigma^2 \\ &= \iint p(y|\mu,\sigma^2)p(\sigma^2|\boldsymbol{y})p(\mu|\sigma^2,\boldsymbol{y})\,d\mu d\sigma^2. \end{aligned}$$

It can be shown that with the normal distribution for y, the normal conditional posterior distribution for μ given σ^2, and the inverse-χ^2 marginal posterior distribution for σ^2, the integral is evaluated to

$$p(y|\boldsymbol{y}) \propto \left(1+\frac{50(y-\bar{y})^2}{49s^2(1+1/50)}\right)^{-25}$$

which is a t-distribution on 49 degrees of freedom centered at $\bar{y}$ and with a scale parameter of $\sqrt{s^2(1+1/n)}$. (In our example, $\bar{y} = 110$, $s^2 = 100$, and $n = 50$.)

3. Application to Multilevel Models

3.1. Monte Carlo Methods

The examples above illustrate how statistical summaries of scientific interest can be expressed as integrals of the posterior distribution. Although in simple cases these integrals can sometimes be computed analytically, in more complex realistic examples, numerical methods are required. Even computing a 95% central posterior interval for the probability of breast cancer, θ, in **Example 1** required numerical methods. In this section, we describe Monte Carlo methods, which have revolutionized applied Bayesian data analysis over the past 20 years. Monte Carlo methods are so important because they are often relatively easy to understand and implement, yet are powerful enough to enable us to compute relevant statistical summaries even when fitting highly structured models.

As an introduction to Monte Carlo methods, we return to the LDL cholesterol study.

Example 2 (Continued)

Monte Carlo methods are simulation-based methods. With a specified probability distribution, a typical Monte Carlo simulation involves a computer program generating multiple plausible values from the distribution. In Bayesian data analysis, this generally involves acquiring a sample from the posterior (or posterior predictive) distribution. In **Figure 3**, we compare a Monte Carlo sample from the posterior distribution with the three plots of the posterior distribution given in **Figure 2**. The key here is that we can draw the same inferences regarding μ and σ^2 from either the plots of the Monte Carlo sample or from the plots of the posterior distribution itself. In addition to the qualitative descriptions discussed in **Section 2.3**, we can compute posterior means by averaging over the Monte Carlo sample or compute a 95% central interval, by computing the 2.5%-ile and 97.5%-ile of the Monte Carlo sample.

Example 2 is a simple illustration with only two parameters. This makes it easy to visually examine the joint posterior distribution and to compute the marginal posterior distributions of the parameters of interest. In more complex settings, however, the dimension of the unknown parameter may be much larger. In image analysis (e.g., functional magnetic resonance imaging), for example, there may be an unknown image intensity in each of a large number of pixels or voxels. In such settings, there may be hundreds or thousands of unknown parameters. It is in such settings that Monte Carlo methods are so

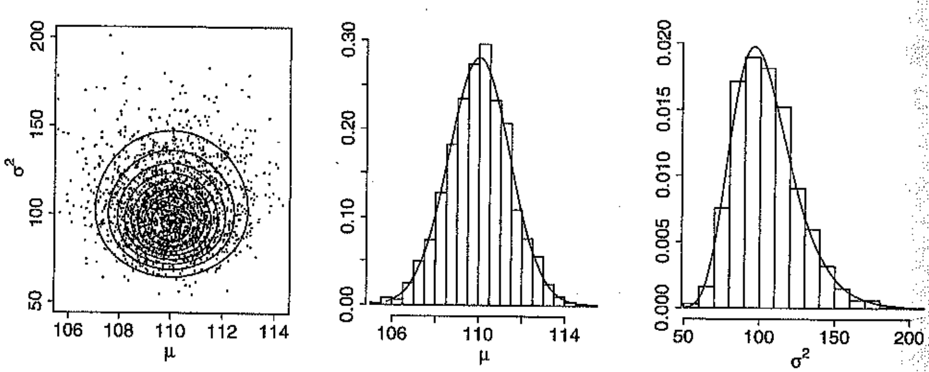

Fig. 3. A Monte Carlo sample from the posterior distribution of parameters of LDL cholesterol levels. A Monte Carlo sample is compared with each of the 3 plots given in Figure 2. The Monte Carlo sample carries the same information about the posterior distribution as the analytically computed plots.

useful. Although we cannot plot the joint posterior distribution or even compute the high-dimensional integrations that are required to evaluate the marginal posterior distributions of low-dimensional quantities of scientific interest, we may be able to acquire a Monte Carlo sample from the posterior distribution. That is, although we cannot produce plots analogous to those in **Figure 2**, we can produce scatter plots and histograms analogous to those in **Figure 3**. From these representations of the Monte Carlo sample, we can construct statistical inferences for unknown quantities of scientific interest, even in highly complex models. This strategy is illustrated in a more complex setting in **Section 3.2**.

There are a variety of techniques available for acquiring a Monte Carlo sample from a given posterior distribution. Perhaps the most important class of such techniques is known as *Markov chain Monte Carlo* (MCMC). It was the development of MCMC in the statistical literature, starting in the late 1980s, that greatly expanded the class of models that can be fit using Monte Carlo techniques. An important example of MCMC is the *Gibbs sampler*. Rather than directly acquiring a Monte Carlo sample from the posterior distribution, the Gibbs sampler cycles through a set of conditional posterior distributions, sampling from each distribution conditional on the most recent draw of the remaining parameters. Because the conditional distributions involve a smaller number of unknown parameters, they tend to be simpler to simulate. Carefully designed Gibbs samplers allow highly complex models to be divided into a sequence of simpler more standard models, all of which can be fit using standard Bayesian statistical techniques. The iterative nature of the Gibbs sampler (and other MCMC techniques) means that it can be sensitive to starting values, and its Monte Carlo nature means that convergence diagnostics can be subtle. Here, we have only scratched the surface of the numerous technical issues involved in designing, implementing, and detecting convergence of MCMC samplers. Nonetheless, interpreting the scientific results is done in much the same way as with the Monte Carlo methods described here. Readers interested in learning more about this important class of Bayesian computational methods are directed to the references in **Section 4** and the citations therein.

3.2. Multilevel Models

The power of Monte Carlo sampling in conjunction with Bayesian methodology is that it allows us to fit models that are explicitly designed to capture the complexity of any given data generation mechanism. We often accomplish this by hierarchically combining a series of simple models into a single more appropriate model. In this section, we illustrate this strategy in an extended example. Although this example is relatively simple by current standards, we hope that it will give the reader a flavor for how multilevel models are constructed and for the power of combining Monte Carlo sampling with Bayesian methods.

Table 1
Data for the 16 Litters of Rats in the Treatment Group

	Litter															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Size	12	11	10	9	11	10	10	9	9	5	9	7	10	6	10	7
Surviving	12	11	10	9	10	9	9	8	8	4	7	4	5	3	3	0

The litter sizes and the number of pups surviving the 21-day lactation period are recorded.

Example 3

In an experiment described by Weil *(2)*, 32 pregnant female rats were divided into 2 groups. In the control group, the mothers were fed a control diet during pregnancy and lactation. In the second group, the mothers' diets were treated with a chemical. The number of pups in each litter that survived 4 days was recorded as the litter size. Of these, the number that survived the 21-day lactation period were also recorded. For our purposes, we consider only the treatment group and investigate how the probability of 21-day survival varies among the litters in this population and fit the probability of survival for each of the 16 observed treatment litters. The data for the treatment litters appear in **Table 1**, which records the size of each litter (number of pups that survive for 4 days) and number of these that survive for 21 days.

We begin by formulating a probability model for the data. For each litter, let n_i be the size of the litter and y_i be the number of pups that survive the 21-day lactation period. We assume the pups within each litter have equal probability of survival and use a binomial distribution to model the number that survive. In particular, we assume $y_i | \theta_i \sim \text{Binomial}(n_i, \theta_i)$, that is,

$$p(y_i|\theta_i) = \binom{n_i}{y_i}\theta_i^{y_i}(1-\theta_i)^{(n_i-y_i)}.$$

Because we believe the survival rates vary among the litters, we allow θ_i to depend on i. The distribution of the θ_i is of primary interest in this study (in particular, we may be interested in how the distribution is affected by the treatment). Therefore we introduce a probability model for the θ_i. As discussed in **Example 1**, the Beta distribution is particularly well suited for modeling probabilities. Thus, we assume $\theta_i \sim \text{Beta}(\alpha, \beta)$. The parameters α and β determine the shape, mean, and variability of the Beta distribution and thus of the survival probabilities among litters in the treatment group.

In **Example 1**, we used prior information as to the probability of breast cancer to set the values of α and β. In this case, however, α and β are fit to the data to describe the distribution of the survival probabilities. Because α and β, both restricted to be positive, are treated as model parameters, we must decide on prior distributions for these 2 parameters. Here we choose independent noninformative prior distributions that are uniform on $\log(\alpha)$ and $\log(\beta)$. As in **Example 2**, this corresponds with prior distributions that are proportional to the reciprocal, that is, $p(\alpha, \beta) \propto 1/\alpha\beta$.

Combining the two parts of the specification of the data model with the prior distribution leads to a 3-level model. In particular, the statistical model can be formulated as a Beta-binomial model *(3)* with noninformative prior distribution as follows:

Level 1: $y_i \mid \theta_i \sim \text{Binomial}(n_i, \theta_i)$ for $i = 1, \ldots, 16$.
Level 2: $\theta_i \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$ for $i = 1, \ldots, 16$.
Level 3: $p(\alpha, \beta) \propto 1/\alpha\beta$.

Level 1 specifies the 16 within-litter distributions, **level 2** describes the variability among the litters in the treatment population, and **level 3** is the (noninformative and improper) prior distribution. This is a simple illustration of how standard probability distributions can be combined hierarchically to form more complex and more appropriate models—models that can more fully describe the richness of the data generation mechanism.

With the data model, prior distribution, and observed data in hand, we construct and compute the posterior distribution as described earlier. We acquire a Monte Carlo sample from the joint posterior distribution of $(\theta_1, \ldots, \theta_{16}, \alpha, \beta)$. **Figure 4** represents the Monte Carlo sample from the marginal posterior

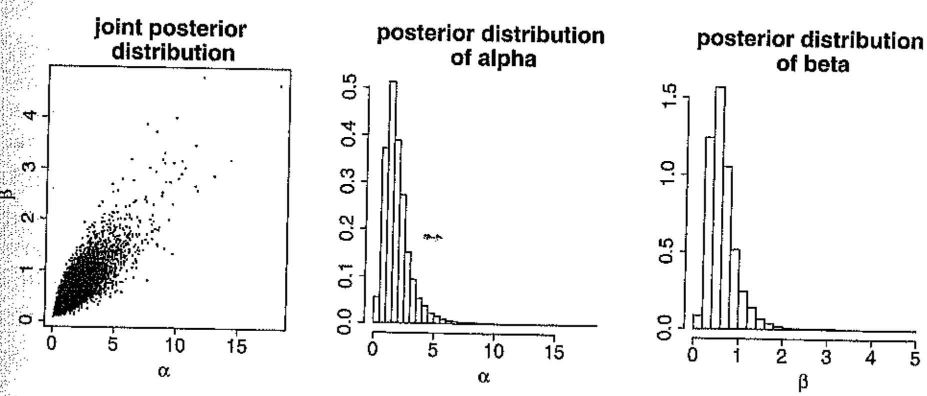

Fig. 4. A Monte Carlo sample from the joint posterior distribution of α and β.

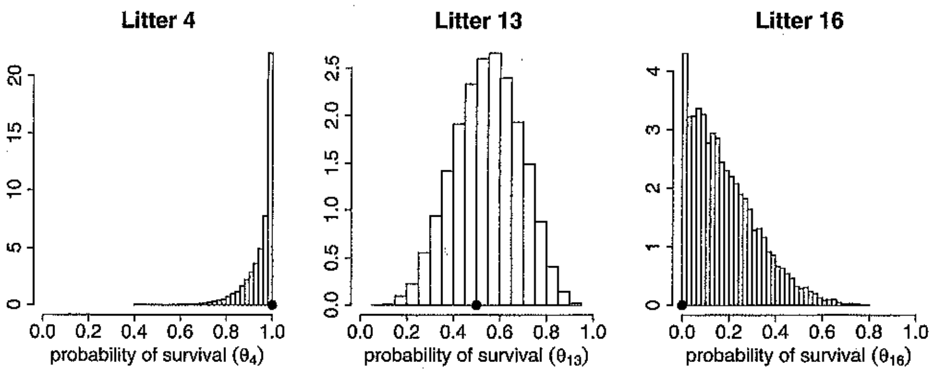

Fig. 5. Histograms of the Monte Carlo sample of the survival probabilities of 3 of the litters. The solid circles on the horizontal axis of each of the histograms represent the sample proportion of the pups that survived in that litter. Notice that in all 3 cases the histograms have their centers of mass a bit off of the sample proportion, in the direction of the fitted population mean of 0.74. This is known as *shrinkage*: the posterior mean "shrinks" from the sample proportion toward the fitted population mean.

distribution of α and β, and **Figure 5** represents a sample from the marginal posterior distributions of θ_4, θ_{13}, and θ_{16}. In this case, the plots in **Figure 5** are more relevant because the parameters are more easily interpreted: they are the marginal posterior distributions of the survival probabilities for 3 of the litters.

Comparing the three plots in **Figure 5**, it is clear that the survival probabilities vary among the litters. To explore this further, we can acquire a Monte Carlo sample from the predictive distribution of the survival probability of another litter. A histogram of this Monte Carlo sample appears in the first panel of **Figure 6**. This distribution accounts for both the variability among the litters and the uncertainty in the distribution of the survival probabilities. These two variance components correspond with the variability among the histograms in **Figure 5** and the uncertainty in α and β illustrated in **Figure 4**, respectively. The final histogram in **Figure 6** is a Monte Carlo sample from the posterior predictive distribution of the number of surviving pups for an additional litter of size 10. This distribution accounts for both the variability in θ as represented by the first histogram in **Figure 6** and for the binomial variation of pup survival.

We can fit the survival probabilities of each of the 16 litters by averaging over the Monte Carlo sample of each of these 16 parameters. The results, along

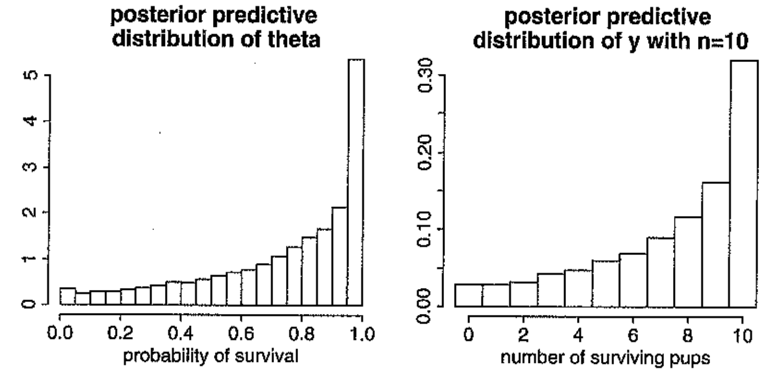

Fig. 6. Monte Carlo samples from the posterior predictive distribution. The first histogram represents a sample from the predictive distribution of the survival probability of another litter from this population. The second histogram corresponds with a sample from the predictive distribution of the number of surviving pups from this additional litter, given that the litter is of size 10.

with the sample proportion of surviving pups in each litter, appear in **Table 2**. Notice that in each case, the fitted probability is between the sample proportion and the expected survival probability of a new litter, 0.74. Although the sample proportion is the standard estimate of the survival probability for a single litter, like all statistical estimates, these have error because of the variable nature of binomial data. Because we are simultaneously fitting the population distribution of survival probabilities, we have some information as to the direction of the estimates' error. The Bayesian estimate is an average of the population mean and the sample proportion. As the size of the litter increases, this average is weighted more heavily toward the sample proportion. These fitted values are often called *shrinkage estimates* because they "shrink" the fitted probability from the sample proportion toward the population mean. Shrinkage is automatic

Table 2
Shrinkage

	Litter															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Sample	1.00	1.00	1.00	1.00	0.91	0.90	0.90	0.89	0.89	0.80	0.78	0.57	0.50	0.50	0.30	0.00
Fitted	0.96	0.96	0.95	0.95	0.88	0.87	0.87	0.86	0.86	0.78	0.77	0.61	0.55	0.57	0.38	0.18

The sample proportion of surviving pups and the fitted probability of survival are recored for each of the 16 litters. Each of the fitted values is between the population mean (0.74) and the sample proportions for the particular litter.

when the Bayesian posterior distribution is used to generate statistical estimates.

4. Other Resources

In this chapter, we have introduced only the most basic aspects of Bayesian modeling, methods, and computation. There are a number of accessible treatises on Bayesian methods that interested readers might refer to, including Gelman and others *(4)* and Carlin and Louis *(5)*, both of whom offer excellent introductions.

References

1. Abramowitz, M., and Stegun, I. A. (1972) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Table*, 9th ed. New York, Dover Publications.
2. Weil, C. S. (1970) Selection of the valid number of sampling units and a consdiera-tion of their combination in toxicological studies involving reproduction, teratongenesis or carcinogensis. *Food and Cosmetics Toxicology* **8**, 177–182.
3. Williams, D. A. (1975) 394: The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**, 949–952.
4. Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003) *Bayesian Data Analysis*, 2nd ed. London, Chapman & Hall.
5. Carlin, B. P., and Louis, T. A. (2000) *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. Boca Raton, Chapman & Hall.

17

Overview of Missing Data Techniques

Ralph B. D'Agostino, Jr.

Summary

Missing data frequently arise in the course of research studies. Understanding the mechanism that led to the missing data is important in order for investigators to be able to perform analyses that will lead to proper inference. This chapter will review different missing data mechanisms, including random and non-random mechanisms. Basic methods will be presented using examples to illustrate approaches to analyzing data in the presence of missing data.

Key Words: Imputation, missing data mechanism, MAR, MCAR, nonignorable missing data

1. Introduction

In the previous 16 chapters, you have been presented with a variety of methods and techniques for analyzing data in order to make valid inference. In these previous chapters, a common assumption concerning the validity of the techniques has been that there is complete data available on all units measured in the experiment or study. The goal of this chapter is to present an overview of what can be done when this assumption is violated and missing data occurs on observations in an experiment or study.

Missing data frequently arise in the course of research studies. This phenomenon, though rarely intended, can have varying impact on the ability of investigators to draw proper conclusions concerning the relevance of their data. Often, the existence of missing data itself is not the issue that is of most importance, but rather understanding the mechanism that led to data being missing is most relevant. If one can understand the mechanisms that led to data being missing, then often appropriate analytical strategies can be used to handle its occurrence. This chapter will introduce basic concepts concerning approaches

From: *Methods in Molecular Biology, vol. 404: Topics in Biostatistics*
Edited by: W. T. Ambrosius © Humana Press Inc., Totowa, NJ