CHAPTER 1

Background

1.1 Overview

By Bayesian data analysis, we mean practical methods for making inferences from data using probability models for quantities we observe and for quantities about which we wish to learn. The essential characteristic of Bayesian methods is their explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis.

The process of Bayesian data analysis can be idealized by dividing it into the following three steps:

1. Setting up a *full probability model*—a joint probability distribution for all observable and unobservable quantities in a problem. The model should be consistent with knowledge about the underlying scientific problem and the data collection process.
2. Conditioning on observed data: calculating and interpreting the appropriate *posterior distribution*—the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.
3. Evaluating the fit of the model and the implications of the resulting posterior distribution: does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions in step 1? If necessary, one can alter or expand the model and repeat the three steps.

Great advances in all these areas have been made in the last forty years, and many of these are reviewed and used in examples throughout the book. Our treatment covers all three steps, the second involving computational methodology and the third a delicate balance of technique and judgment, guided by the applied context of the problem. The first step remains a major stumbling block for much Bayesian analysis: just where do our models come from? How do we go about constructing appropriate probability specifications? We provide some guidance on these issues and illustrate the importance of the third step in retrospectively evaluating the fit of models. Along with the improved techniques available for computing conditional probability distributions in the second step, advances in carrying out the third step alleviate to some degree the need for completely correct model specification at the first attempt. In particular, the much-feared dependence of conclusions on 'subjective' prior distributions can be examined and explored.

A primary motivation for believing Bayesian thinking important is that it facilitates a common-sense interpretation of statistical conclusions. For instance,

a Bayesian (probability) interval for an unknown quantity of interest can be directly regarded as having a high probability of containing the unknown quantity, in contrast to a frequentist (confidence) interval, which may strictly be interpreted only in relation to a sequence of similar inferences that might be made in repeated practice. Recently in applied statistics, increased emphasis has been placed on interval estimation rather than hypothesis testing, and this provides a strong impetus to the Bayesian viewpoint, since it seems likely that most users of standard confidence intervals give them a common-sense Bayesian interpretation. One of our aims in this book is to indicate the extent to which Bayesian interpretations of common simple statistical procedures are justified.

Rather than engage in philosophical debates about the foundations of statistics, however, we prefer to concentrate on the pragmatic advantages of the Bayesian framework, whose flexibility and generality allow it to cope with very complex problems. The central feature of Bayesian inference, the direct quantification of uncertainty, means that there is no impediment in principle to fitting models with many parameters and complicated multilayered probability specifications. In practice, the problems are ones of setting up and computing with large models, and a large part of this book focuses on recently developed and still developing techniques for handling these modeling and computational challenges. The freedom to set up complex models arises in large part from the fact that the Bayesian paradigm provides a conceptually simple method for coping with multiple parameters, as we discuss in detail from Chapter 3 on.

1.2 General notation for statistical inference

Statistical inference is concerned with drawing conclusions, from numerical data, about quantities that are not observed. For example, a clinical trial of a new cancer drug might be designed to compare the five-year survival probability in a population given the new drug with that in a population under standard treatment. These survival probabilities refer to a large *population* of patients, and it is neither feasible nor ethically acceptable to experiment with an entire population. Therefore inferences about the true probabilities and, in particular, their differences must be based on a *sample* of patients. In this example, even if it were possible to expose the entire population to one or the other treatment, it is obviously never possible to expose anyone to both treatments, and therefore statistical inference would still be needed to assess the *causal inference*—the comparison between the observed outcome in each patient and that patient's unobserved outcome if exposed to the other treatment.

We distinguish between two kinds of *estimands*—unobserved quantities for which statistical inferences are made—first, potentially observable quantities, such as future observations of a process, or the outcome under the treatment not received in the clinical trial example; and second, quantities that are not

directly observable, that is, parameters that govern the hypothetical process leading to the observed data (for example, regression coefficients). The distinction between these two kinds of estimands is not always precise, but is generally useful as a way of understanding how a statistical model for a particular problem fits into the real world.

Parameters, data, and predictions

As general notation, we let θ denote unobservable vector quantities or population *parameters* of interest (such as the probabilities of survival under each treatment for randomly chosen members of the population in the example of the clinical trial), y denote the observed data (such as the numbers of survivors and deaths in each treatment group), and $\tilde{y}$ denote unknown, but potentially observable, quantities (such as the outcomes of the patients under the other treatment, or the outcome under each of the treatments for a new patient similar to those already in the trial). In general these symbols represent multivariate quantities. We generally use Greek letters for parameters, lower-case Roman letters for observed or observable scalars and vectors (and sometimes matrices), and upper-case Roman letters for observed or observable matrices. When using matrix notation, we consider vectors as column vectors throughout; for example, if u is a vector with n components, then $u^T u$ is a scalar and uu^T an $n \times n$ matrix.

Observational units and variables

In many statistical studies, data are gathered on each of a set of n objects or *units*, and we can write the data as a vector, $y = (y_1, \ldots, y_n)$. In the clinical trial example, we might label y_i as 1 if patient i is alive after five years or 0 if the patient dies. If several variables are measured on each unit, then each y_i is actually a vector, and the entire dataset y is a matrix (usually taken to have n rows). The y variables are called the 'outcomes' and are considered 'random' in the sense that, when making inferences, we wish to allow for the possibility that the observed values of the variables could have turned out otherwise, due to the sampling process and the natural variation of the population.

Exchangeability

The usual starting point of a statistical analysis is the (often tacit) assumption that the n values y_i may be regarded as *exchangeable*, meaning that the joint probability density $p(y_1, \ldots, y_n)$ should be invariant to permutations of the indexes. A nonexchangeable model would be appropriate if information relevant to the outcome were conveyed in the unit indexes rather than by explanatory variables (see below). The idea of exchangeability is fundamental to statistics, and we return to it repeatedly throughout the book.

Generally, it is useful and appropriate to model data from an exchangeable

distribution as independently and identically distributed (*iid*) given some unknown parameter vector θ with distribution $p(\theta)$. In the clinical trial example, we might model the outcomes y_i as iid, given θ, the unknown probability of survival.

Explanatory variables

It is common to have observations on each unit that we do not bother to model as random. In the clinical trial example, such variables might include the age and previous health status of each patient in the study. We call this second class of variables *explanatory variables*, or *covariates*, and label them x. We use X to denote the entire set of explanatory variables for all n units; if there are k explanatory variables, then X is a matrix with n rows and k columns. Treating X as random, the notion of exchangeability can be extended to require the distribution of the n values of $(x, y)_i$ to be unchanged by arbitrary permutations of the indexes. It is *always* appropriate to assume an exchangeable model after incorporating sufficient relevant information in X that the indexes can be thought of as randomly assigned. It follows from the assumption of exchangeability that the distribution of y, given x, is the same for all units in the study in the sense that if two units have the same value of x, then their distributions of y are the same. Any of the explanatory variables x can of course be moved into the y category if we wish to model them. We discuss the role of explanatory variables (also called predictors) in detail in Chapter 7 in the context of analyzing surveys, experiments, and observational studies, and in Parts IV and V in the context of regression models.

Hierarchical modeling

In Chapter 5 and subsequent chapters, we focus on *hierarchical models* (also called *multilevel models*), which are used when information is available on several different levels of observational units. In a hierarchical model, it is possible to speak of exchangeability at each level of units. For example, suppose two medical treatments are applied, in separate randomized experiments, to patients in several different cities. Then, if no other information were available, it would be reasonable to treat the patients within each city as exchangeable and also treat the results from different cities as themselves exchangeable. In practice it would make sense to include, as explanatory variables at the city level, whatever relevant information we have on each city, as well as the explanatory variables mentioned before at the individual level, and then the conditional distributions given these explanatory variables would be exchangeable.

1.3 Bayesian inference

Bayesian statistical conclusions about a parameter θ, or unobserved data $\tilde{y}$, are made in terms of *probability* statements. These probability statements are

conditional on the observed value of y, and in our notation are written simply as $p(\theta|y)$ or $p(\tilde{y}|y)$. We also implicitly condition on the known values of any covariates, x. It is at the fundamental level of conditioning on observed data that Bayesian inference departs from the approach to statistical inference described in many textbooks, which is based on a retrospective evaluation of the procedure used to estimate θ (or $\tilde{y}$) over the distribution of possible y values conditional on the true unknown value of θ. Despite this difference, it will be seen that in many simple analyses, superficially similar conclusions result from the two approaches to statistical inference. However, analyses obtained using Bayesian methods can be easily extended to more complex problems. In this section, we present the basic mathematics and notation of Bayesian inference, followed in the next section by an example from genetics.

Probability notation

Some comments on notation are needed at this point. First, $p(\cdot|\cdot)$ denotes a conditional probability density with the arguments determined by the context, and similarly for $p(\cdot)$, which denotes a marginal distribution. We use the terms 'distribution' and 'density' interchangeably. The same notation is used for continuous density functions and discrete probability mass functions. Different distributions in the same equation (or expression) will each be denoted by $p(\cdot)$, as in (1.1) below, for example. Although an abuse of standard mathematical notation, this method is compact and similar to the standard practice of using $p(\cdot)$ for the probability of any discrete event, where the sample space is also suppressed in the notation. Depending on context, to avoid confusion, we may use the notation $\Pr(\cdot)$ for the probability of an event; for example, $\Pr(\theta > 2) = \int_{\theta>2} p(\theta)d\theta$. When using a standard distribution, we use a notation based on the name of the distribution; for example, if θ has a normal distribution with mean μ and variance σ^2, we write $\theta \sim \mathrm{N}(\mu, \sigma^2)$ or $p(\theta) = \mathrm{N}(\theta|\mu, \sigma^2)$ or, to be even more explicit, $p(\theta|\mu, \sigma^2) = \mathrm{N}(\theta|\mu, \sigma^2)$. Throughout, we use notation such as $\mathrm{N}(\mu, \sigma^2)$ for random variables and $\mathrm{N}(\theta|\mu, \sigma^2)$ for density functions. Notation and formulas for several standard distributions appear in Appendix A.

We also occasionally use the following expressions for all-positive random variables θ: the *coefficient of variation* (CV) is defined as $\mathrm{sd}(\theta)/\mathrm{E}(\theta)$, the *geometric mean* (GM) is $\exp(\mathrm{E}[\log(\theta)])$, and the geometric standard deviation (GSD) is $\exp(\mathrm{sd}[\log(\theta)])$.

Bayes' rule

In order to make probability statements about θ given y, we must begin with a *model* providing a *joint probability distribution* for θ and y. The joint probability mass or density function can be written as a product of two densities that are often referred to as the *prior distribution* $p(\theta)$ and the *sampling distribution* (or *data distribution*) $p(y|\theta)$ respectively:

$$p(\theta, y) = p(\theta)p(y|\theta).$$

Simply conditioning on the known value of the data y, using the basic property of conditional probability known as Bayes' rule, yields the *posterior* density:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}, \tag{1.1}$$

where $p(y) = \sum_\theta p(\theta)p(y|\theta)$, and the sum is over all possible values of θ (or $p(y) = \int p(\theta)p(y|\theta)d\theta$ in the case of continuous θ). An equivalent form of (1.1) omits the factor $p(y)$, which does not depend on θ and, with fixed y, can thus be considered a constant, yielding the *unnormalized posterior density* , which is the right side of (1.2):

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{1.2}$$

These simple expressions encapsulate the technical core of Bayesian inference: the primary task of any specific application is to develop the model $p(\theta, y)$ and perform the necessary computations to summarize $p(\theta|y)$ in appropriate ways.

Prediction

To make inferences about an unknown observable, often called predictive inferences, we follow a similar logic. Before the data y are considered, the distribution of the unknown but observable y is

$$p(y) = \int p(y, \theta)d\theta = \int p(\theta)p(y|\theta)d\theta. \tag{1.3}$$

This is often called the marginal distribution of y, but a more informative name is the *prior predictive distribution*: prior because it is not conditional on a previous observation of the process, and predictive because it is the distribution for a quantity that is observable.

After the data y have been observed, we can predict an unknown observable, $\tilde{y}$, from the same process. For example, $y = (y_1, \ldots, y_n)$ may be the vector of recorded weights of an object weighed n times on a scale, $\theta = (\mu, \sigma^2)$ may be the unknown true weight of the object and the measurement variance of the scale, and $\tilde{y}$ may be the yet to be recorded weight of the object in a planned new weighing. The distribution of $\tilde{y}$ is called the *posterior predictive distribution*, posterior because it is conditional on the observed y and predictive because it is a prediction for an observable $\tilde{y}$:

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y)d\theta \\ &= \int p(\tilde{y}|\theta, y)p(\theta|y)d\theta \\ &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta. \end{aligned} \tag{1.4}$$

The second and third lines display the posterior predictive distribution as an average of conditional predictions over the posterior distribution of θ. The last

equation follows because y and $\tilde{y}$ are conditionally independent given θ in this model.

Likelihood

Using Bayes' rule with a chosen probability model means that the data y affect the posterior inference (1.2) *only* through the function $p(y|\theta)$, which, when regarded as a function of θ, for fixed y, is called the *likelihood function*. In this way Bayesian inference obeys what is sometimes called the *likelihood principle*, which states that for a given sample of data, any two probability models $p(y|\theta)$ that have the same likelihood function yield the same inference for θ.

The likelihood principle is reasonable, but only within the framework of the model or family of models adopted for a particular analysis. In practice, one can rarely be confident that the chosen model is *the* correct model. We shall see in Chapter 6 that sampling distributions (imagining repeated realizations of our data) can play an important role in checking model assumptions. In fact, our view of an applied Bayesian statistician is one who is willing to apply Bayes' rule under a variety of possible models.

Likelihood and odds ratios

The ratio of the posterior density $p(\theta|y)$ evaluated at the points θ_1 and θ_2 under a given model is called the posterior *odds* for θ_1 compared to θ_2. The most familiar application of this concept is with discrete parameters, with θ_2 taken to be the complement of θ_1. Odds provide an alternative representation of probabilities and have the attractive property that Bayes' rule takes a particularly simple form when expressed in terms of them:

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(\theta_1)p(y|\theta_1)/p(y)}{p(\theta_2)p(y|\theta_2)/p(y)} = \frac{p(\theta_1)}{p(\theta_2)}\frac{p(y|\theta_1)}{p(y|\theta_2)}. \tag{1.5}$$

In words, the posterior odds are equal to the prior odds multiplied by the *likelihood ratio*, $p(y|\theta_1)/p(y|\theta_2)$.

1.4 Example: inference about a genetic probability

The following example is not typical of *statistical* applications of the Bayesian method, because it deals with a very small amount of data and concerns a single individual's state (gene carrier or not) rather than with the estimation of a parameter that describes an entire population. Nevertheless it is a real example of the very simplest type of Bayesian calculation, where the estimand and the individual item of data each have only two possible values.

Human males have one X-chromosome and one Y-chromosome, whereas females have two X-chromosomes, each chromosome being inherited from one parent. Hemophilia is a disease that exhibits X-chromosome-linked recessive

inheritance, meaning that a male who inherits the gene that causes the disease on the X-chromosome is affected, whereas a female carrying the gene on only one of her two X-chromosomes is not affected. The disease is generally fatal for women who inherit two such genes, and this is very rare, since the frequency of occurrence of the gene is low in human populations.

The prior distribution

Consider a woman who has an affected brother, which implies that her mother must be a carrier of the hemophilia gene with one 'good' and one 'bad' hemophilia gene. We are also told that her father is not affected; thus the woman herself has a fifty-fifty chance of having the gene. The unknown quantity of interest, the state of the woman, has just two values: the woman is either a carrier of the gene ($\theta = 1$) or not ($\theta = 0$). Based on the information provided thus far, the prior distribution for the unknown θ can be expressed simply as $\Pr(\theta = 1) = \Pr(\theta = 0) = \frac{1}{2}$.

The model and likelihood

The data used to update this prior information consist of the affection status of the woman's sons. Suppose she has two sons, neither of whom is affected. Let $y_i = 1$ or 0 denote an affected or unaffected son, respectively. The outcomes of the two sons are exchangeable and, conditional on the unknown θ, are independent; we assume the sons are not identical twins. The two items of independent data generate the following likelihood function:

$$\begin{aligned} \Pr(y_1 = 0, y_2 = 0 \,|\, \theta = 1) &= (0.5)(0.5) = 0.25 \\ \Pr(y_1 = 0, y_2 = 0 \,|\, \theta = 0) &= (1)(1) = 1. \end{aligned}$$

These expressions follow from the fact that if the woman is a carrier, then each of her sons will have a 50% chance of inheriting the gene and so being affected, whereas if she is not a carrier then there is a probability very close to 1 that a son of hers will be unaffected. (In fact, there is a nonzero probability of being affected even if the mother is not a carrier, but this risk—the mutation rate—is very small and can be ignored for this example.)

The posterior distribution

Bayes' rule can now be used to combine the information in the data with the prior probability; in particular, interest is likely to focus on the posterior probability that the woman is a carrier. Using y to denote the joint data (y_1, y_2), this is simply

$$\begin{aligned} \Pr(\theta = 1|y) &= \frac{p(y|\theta = 1)\Pr(\theta = 1)}{p(y|\theta = 1)\Pr(\theta = 1) + p(y|\theta = 0)\Pr(\theta = 0)} \\ &= \frac{(0.25)(0.5)}{(0.25)(0.5) + (1.0)(0.5)} = \frac{0.125}{0.625} = 0.20. \end{aligned}$$

Intuitively it is clear that if a woman has unaffected children, it is less probable that she is a carrier, and Bayes' rule provides a formal mechanism for determining the extent of the correction. The results can also be described in terms of prior and posterior odds. The prior odds of the woman being a carrier are $0.5/0.5 = 1$. The likelihood ratio based on the information about her two unaffected sons is $0.25/1 = 0.25$, so the posterior odds are obtained very simply as 0.25. Converting back to a probability, we obtain $0.25/(1+0.25) = 0.2$, just as before.

Adding more data

A key aspect of Bayesian analysis is the ease with which sequential analyses can be performed. For example, suppose that the woman has a third son, who is also unaffected. The entire calculation does not need to be redone; rather we use the previous posterior distribution as the new prior distribution, to obtain:

$$\Pr(\theta = 1|y_1, y_2, y_3) = \frac{(0.5)(0.20)}{(0.5)(0.20) + (1)(0.8)} = 0.111.$$

Alternatively, if we suppose that the third son is affected, it is easy to check that the posterior probability of the woman being a carrier becomes 1 (again ignoring the possibility of a mutation).

1.5 Probability as a measure of uncertainty

We have already used concepts such as probability density, and indeed we assume that the reader has a fair degree of familiarity with basic probability theory (although in Section 1.8 we provide a brief technical review of some probability calculations that often arise in Bayesian analysis). But since the uses of probability within a Bayesian framework are much broader than within non-Bayesian statistics, it is important to consider at least briefly the foundations of the concept of probability before considering more detailed statistical examples. We take for granted a common understanding on the part of the reader of the mathematical definition of probability: that probabilities are numerical quantities, defined on a set of 'outcomes,' that are nonnegative, additive over mutually exclusive outcomes, and sum to 1 over all possible mutually exclusive outcomes.

In Bayesian statistics, probability is used as the fundamental measure or yardstick of uncertainty. Within this paradigm, it is equally legitimate to discuss the probability of 'rain tomorrow' or of a Brazilian victory in the soccer World Cup as it is to discuss the probability that a coin toss will land heads. Hence, it becomes as natural to consider the probability that an unknown estimand lies in a particular range of values as it is to consider the probability that the mean of a random sample of 10 items from a known fixed population of size 100 will lie in a certain range. The first of these two probabilities is of more interest after data have been acquired whereas the second is more

relevant beforehand. Bayesian methods enable statements to be made about the partial knowledge available (based on data) concerning some situation or 'state of nature' (unobservable or as yet unobserved) in a systematic way, using probability as the yardstick. The guiding principle is that the state of knowledge about anything unknown is described by a probability distribution.

What is meant by a numerical measure of uncertainty? For example, the probability of 'heads' in a coin toss is widely agreed to be $\frac{1}{2}$. Why is this so? Two justifications seem to be commonly given:

1. Symmetry or exchangeability argument:

$$\text{probability} = \frac{\text{number of favorable cases}}{\text{number of possibilities}},$$

   assuming equally likely possibilities. For a coin toss this is really a physical argument, based on assumptions about the forces at work in determining the manner in which the coin will fall, as well as the initial physical conditions of the toss.

2. Frequency argument: probability = relative frequency obtained in a very long sequence of tosses, assumed to be performed in an identical manner, physically independently of each other.

Both the above arguments are in a sense subjective, in that they require judgments about the nature of the coin and the tossing procedure, and both involve semantic arguments about the meaning of equally likely events, identical measurements, and independence. The frequency argument may be perceived to have certain special difficulties, in that it involves the hypothetical notion of a very long sequence of identical tosses. If taken strictly, this point of view does not allow a statement of probability for a single coin toss that does not happen to be embedded, at least conceptually, in a long sequence of identical events.

The following examples illustrate how probability judgments can be increasingly subjective. First, consider the following modified coin experiment. Suppose that a particular coin is stated to be either double-headed *or* double-tailed, with no further information provided. Can one still talk of the probability of heads? It seems clear that in common parlance one certainly can. It is less clear, perhaps, how to assess this new probability, but many would agree on the same value of $\frac{1}{2}$, perhaps based on the exchangeability of the labels 'heads' and 'tails.'

Now consider some further examples. Suppose Colombia plays Brazil in soccer tomorrow: what is the probability of Colombia winning? What is the probability of rain tomorrow? What is the probability that Colombia wins, if it rains tomorrow? What is the probability that the next space shuttle launched will explode? Although each of these questions seems reasonable in a commonsense way, it is difficult to contemplate strong frequency interpretations for the probabilities being referenced. Frequency interpretations can usually be *constructed*, however, and this is an extremely useful tool in statistics. For example, we can consider the next space shuttle launch as a sample from

the population of potential space shuttle launches, and look at the frequency of past shuttle launches that have exploded (see the bibliographic note at the end of this chapter for more details on this example). Doing this sort of thing scientifically means creating a probability model (or, at the very least, a 'reference set' of comparable events), and this brings us back to a situation analogous to the simple coin toss, where we must consider the outcomes in question as exchangeable and thus equally likely.

Why is probability a reasonable way of quantifying uncertainty? The following reasons are often advanced.

1. By analogy: physical randomness induces uncertainty, so it seems reasonable to describe uncertainty in the language of random events. Common speech uses many terms such as 'probably' and 'unlikely,' and it appears consistent with such usage to extend a more formal probability calculus to problems of scientific inference.
2. Axiomatic or normative approach: related to decision theory, this approach places all statistical inference in the context of decision-making with gains and losses. Then reasonable axioms (ordering, transitivity, and so on) imply that uncertainty *must* be represented in terms of probability. We view this normative rationale as suggestive but not compelling.
3. Coherence of bets. *Define* the probability p attached (by you) to an event E as the fraction ($p \in [0, 1]$) at which you would exchange (that is, bet) $\$p$ for a return of \$1 if E occurs. That is, if E occurs, you gain $\$(1-p)$; if the complement of E occurs, you lose $\$p$. For example:

   - Coin toss: thinking of the coin toss as a fair bet suggests even odds corresponding to $p = \frac{1}{2}$.
   - Odds for a game: if you are willing to bet on team A to win a game at 10 to 1 odds against team B (that is, you bet 1 to win 10), your 'probability' for team A winning is at least 1/11.

   The principle of coherence of probabilities states that your assignment of probabilities to all possible events should be such that it is not possible to make a definite gain by betting with you. It can be proved that probabilities constructed under this principle must satisfy the basic axioms of probability theory.

   The betting rationale has some fundamental difficulties:

   - Exact odds are required, on which you would be willing to bet in either direction, for all events. How can you assign exact odds if you are not sure?
   - If a person is willing to bet with you, and has information you do not, it might not be wise for you to take the bet. In practice, probability is an incomplete (necessary but not sufficient) guide to betting.

All of these considerations suggest that probabilities may be a reasonable approach to summarizing uncertainty in applied statistics, but the ultimate

proof is in the success of the applications. The remaining chapters of this book demonstrate that probability provides a rich and flexible framework for handling uncertainty in statistical applications.

Subjectivity and objectivity

All statistical methods that use probability are subjective in the sense of relying on mathematical idealizations of the world. Bayesian methods are sometimes said to be especially subjective because of their reliance on a prior distribution, but in most problems, scientific judgment is necessary to specify both the 'likelihood' and the 'prior' parts of the model. For example, linear regression models are generally at least as suspect as any prior distribution that might be assumed about the regression parameters. A general principle is at work here: whenever there is replication, in the sense of many exchangeable units observed, there is scope for estimating features of a probability distribution from data and thus making the analysis more 'objective.' If an experiment as a whole is replicated several times, then the parameters of the prior distribution can themselves be estimated from data, as discussed in Chapter 5. In any case, however, certain elements requiring scientific judgment will remain, notably the choice of data included in the analysis, the parametric forms assumed for the distributions, and the ways in which the model is checked.

1.6 Example of probability assignment: football point spreads

As an example of how probabilities might be assigned using empirical data and plausible substantive assumptions, we consider methods of estimating the probabilities of certain outcomes in professional (American) football games. This is an example only of probability assignment, not of Bayesian inference. A number of approaches to assigning probabilities for football game outcomes are illustrated: making subjective assessments, using empirical probabilities based on observed data, and constructing a parametric probability model.

Football point spreads and game outcomes

Football experts provide a *point spread* for every football game as a measure of the difference in ability between the two teams. For example, team A might be a 3.5-point favorite to defeat team B. The implication of this point spread is that the proposition that team A, the favorite, defeats team B, the underdog, by 4 or more points is considered a fair bet; in other words, the probability that A wins by more than 3.5 points is $\frac{1}{2}$. If the point spread is an integer, then the implication is that team A is as likely to win by more points than the point spread as it is to win by fewer points than the point spread (or to lose); there is positive probability that A will win by exactly the point spread, in which case neither side is paid off. The assignment of point spreads is itself an interesting exercise in probabilistic reasoning; one interpretation is that the

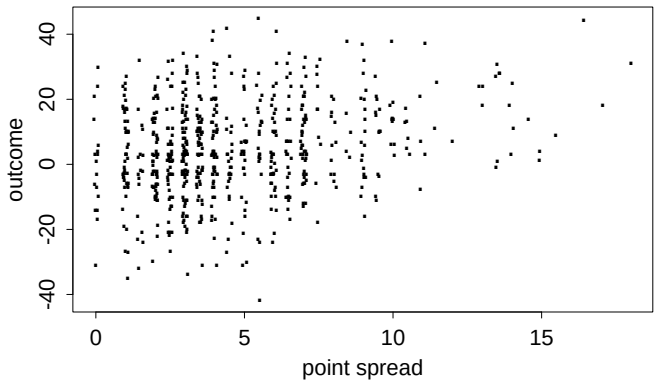

Figure 1.1 *Scatterplot of actual outcome vs. point spread for each of 672 professional football games. The x and y coordinates are jittered by adding uniform random numbers to each point's coordinates (between -0.1 and -0.1 for the x coordinate; between -0.2 and 0.2 for the y coordinate) in order to display multiple values but preserve the discrete-valued nature of each.*

point spread is the median of the distribution of the gambling population's beliefs about the possible outcomes of the game. For the rest of this example, we treat point spreads as given and do not worry about how they were derived.

The point spread and actual game outcome for 672 professional football games played during the 1981, 1983, and 1984 seasons are graphed in Figure 1.1. (Much of the 1982 season was canceled due to a labor dispute.) Each point in the scatterplot displays the point spread, x, and the actual outcome (favorite's score minus underdog's score), y. (In games with a point spread of zero, the labels 'favorite' and 'underdog' were assigned at random.) A small random jitter is added to the x and y coordinate of each point on the graph so that multiple points do not fall exactly on top of each other.

Assigning probabilities based on observed frequencies

It is of interest to assign probabilities to particular events: Pr(favorite wins), Pr(favorite wins | point spread is 3.5 points), Pr(favorite wins by more than the point spread), Pr(favorite wins by more than the point spread | point spread is 3.5 points), and so forth. We might report a subjective probability based on informal experience gathered by reading the newspaper and watching football games. The probability that the favored team wins a game should certainly be greater than 0.5, perhaps between 0.6 and 0.75? More complex events require more intuition or knowledge on our part. A more systematic approach is to assign probabilities based on the data in Figure 1.1. Counting a tied game as one-half win and one-half loss, and ignoring games for which the point spread is zero (and thus there is no favorite), we obtain empirical estimates such as:

- Pr(favorite wins) $= \frac{410.5}{655} = 0.63$

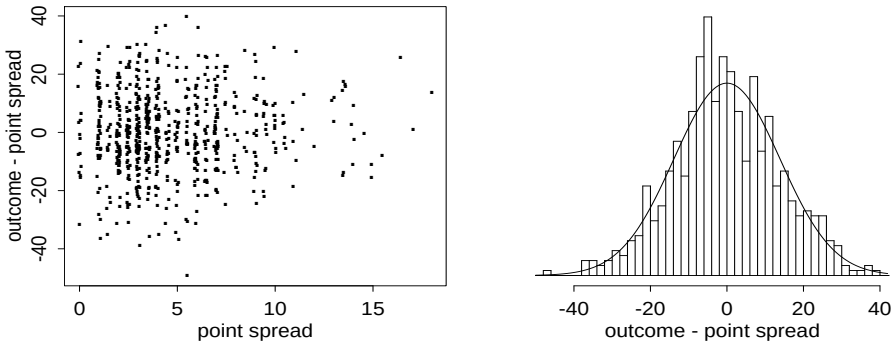

Figure 1.2 *(a) Scatterplot of (actual outcome − point spread) vs. point spread for each of 672 professional football games (with uniform random jitter added to x and y coordinates). (b) Histogram of the differences between the game outcome and the point spread, with the $N(0, 14^2)$ density superimposed.*

- Pr(favorite wins $| \, x = 3.5) = \frac{36}{59} = 0.61$
- Pr(favorite wins by more than the point spread) $= \frac{308}{655} = 0.47$
- Pr(favorite wins by more than the point spread $| \, x = 3.5) = \frac{32}{59} = 0.54$.

These empirical probability assignments all seem sensible in that they match the intuition of knowledgeable football fans. However, such probability assignments are problematic for events with few directly relevant data points. For example, 8.5-point favorites won five out of five times during this three-year period, whereas 9-point favorites won thirteen out of twenty times. However, we realistically expect the probability of winning to be greater for a 9-point favorite than for an 8.5-point favorite. The small sample size with point spread 8.5 leads to very imprecise probability assignments. We consider an alternative method using a parametric model.

A parametric model for the difference between game outcome and point spread

Figure 1.2a displays the differences $y - x$ between the observed game outcome and the point spread, plotted versus the point spread, for the games in the football dataset. (Once again, random jitter was added to both coordinates.) This plot suggests that it may be roughly reasonable to model the distribution of $y - x$ as independent of x. (See Exercise 6.16.) Figure 1.2b is a histogram of the differences $y - x$ for all the football games, with a fitted normal density superimposed. This plot suggests that it may be reasonable to approximate the marginal distribution of the random variable $d = y - x$ by a normal distribution. The sample mean of the 672 values of d is 0.07, and the sample standard deviation is 13.86, suggesting that the results of football games are approximately normal with mean equal to the point spread and standard deviation nearly 14 points (two converted touchdowns). For the remainder of

the discussion we take the distribution of d to be independent of x and normal with mean zero and standard deviation 14 for each x; that is,

$$d|x \sim \mathrm{N}(0, 14^2),$$

as displayed in Figure 1.2b. We return to this example in Sections 2.7 and 3.4 to estimate the parameters of this normal distribution using Bayesian methods. The assigned probability model is not perfect: it does not fit the data exactly, and, as is often the case with real data, neither football scores nor point spreads are continuous-valued quantities.

Assigning probabilities using the parametric model

Nevertheless, the model provides a convenient approximation that can be used to assign probabilities to events. If d has a normal distribution with mean zero and is independent of the point spread, then the probability that the favorite wins by more than the point spread is $\frac{1}{2}$, conditional on any value of the point spread, and therefore unconditionally as well. Denoting probabilities obtained by the normal model as $\Pr_{\mathrm{norm}}$, the probability that an x-point favorite wins the game can be computed, assuming the normal model, as follows:

$$\Pr_{\mathrm{norm}}(y>0\,|\,x) = \Pr_{\mathrm{norm}}(d>-x\,|\,x) = 1 - \Phi\left(-\frac{x}{14}\right),$$

where Φ is the standard normal cumulative distribution function. For example,

- $\Pr_{\mathrm{norm}}(\text{favorite wins}\,|\,x = 3.5) = 0.60$
- $\Pr_{\mathrm{norm}}(\text{favorite wins}\,|\,x = 8.5) = 0.73$
- $\Pr_{\mathrm{norm}}(\text{favorite wins}\,|\,x = 9.0) = 0.74.$

The probability for a 3.5-point favorite agrees with the empirical value given earlier, whereas the probabilities for 8.5- and 9-point favorites make more intuitive sense than the empirical values based on small samples.

1.7 Example of probability assignment: estimating the accuracy of record linkage

We emphasize the essentially empirical (not 'subjective' or 'personal') nature of probabilities with another example in which they are estimated from data.

Record linkage refers to the use of an algorithmic technique to identify records from different databases that correspond to the same individual. Record-linkage techniques are used in a variety of settings. The work described here was formulated and first applied in the context of record linkage between the U.S. Census and a large-scale post-enumeration survey, which is the first step of an extensive matching operation conducted to evaluate census coverage for subgroups of the population. The goal of this first step is to declare as many records as possible 'matched' by computer without an excessive rate of error, thereby avoiding the cost of the resulting manual processing for all records not declared 'matched.'

Existing methods for assigning scores to potential matches

Much attention has been paid in the record-linkage literature to the problem of assigning 'weights' to individual fields of information in a multivariate record and obtaining a composite 'score,' which we call y, that summarizes the closeness of agreement between two records. Here, we assume that this step is complete in the sense that these rules have been chosen. The next step is the assignment of candidate matched pairs, where each pair of records consists of the best potential match for each other from the respective data bases. The specified weighting rules then order the candidate matched pairs. In the motivating problem at the Census Bureau, a binary choice is made between the alternatives 'declare matched' vs. 'send to followup,' where a cutoff score is needed above which records are declared matched. The false-match rate is then defined as the number of falsely matched pairs divided by the number of declared matched pairs.

Particularly relevant for any such decision problem is an accurate method for assessing the probability that a candidate matched pair is a correct match as a function of its score. Simple methods exist for converting the scores into probabilities, but these lead to extremely inaccurate, typically grossly optimistic, estimates of false-match rates. For example, a manual check of a set of records with nominal false-match probabilities ranging from 10^{-3} to 10^{-7} (that is, pairs deemed almost certain to be matches) found actual false-match rates closer to the 1% range. Records with nominal false-match probabilities of 1% had an actual false-match rate of 5%.

We would like to use Bayesian methods to recalibrate these to obtain objective probabilities of matching for a given decision rule—in the same way that in the football example, we used past data to estimate the probabilities of different game outcomes conditional on the point spread. Our approach is to work with the scores y and empirically estimate the probability of a match as a function of y.

Estimating match probabilities empirically

We obtain accurate match probabilities using mixture modeling, a topic we discuss in detail in Chapter 18. The distribution of previously-obtained scores for the candidate matches is considered a 'mixture' of a distribution of scores for true matches and a distribution for non-matches. The parameters of the mixture model are estimated from the data. The estimated parameters allow us to calculate an estimate of the probability of a false match (a pair declared matched that is not a true match) for any given decision threshold on the scores. In the procedure that was actually used, some elements of the mixture model (for example, the optimal transformation required to allow a mixture of normal distributions to apply) were fit using 'training' data with known match status (separate from the data to which we apply our calibration procedure), but we do not describe those details here. Instead we focus on how the method would be used with a set of data with unknown match status.

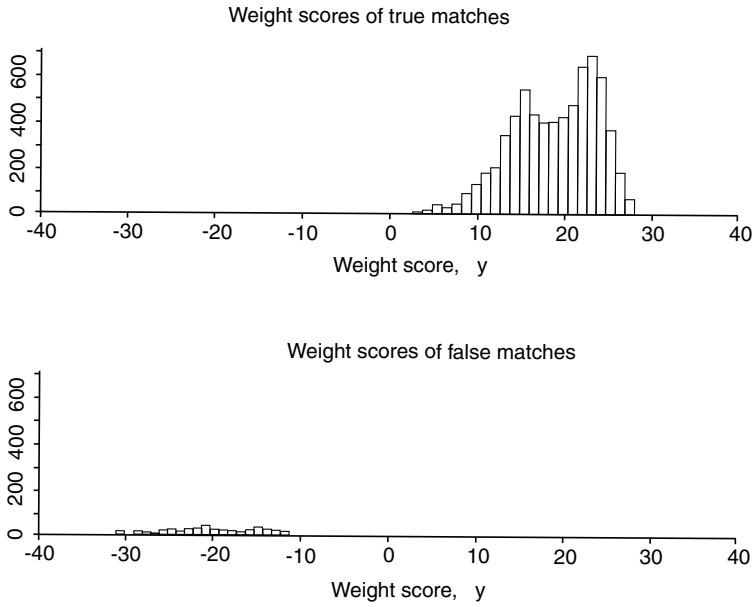

Figure 1.3 *Histograms of weight scores y for true and false matches in a sample of records from the 1988 test Census. Most of the matches in the sample are true (because a pre-screening process has already picked these as the best potential match for each case), and the two distributions are mostly, but not completely, separated.*

Support for this approach is provided in Figure 1.3, which displays the distribution of scores for the matches and non-matches in a particular data set obtained from 2300 records from a 'test Census' survey conducted in a single local area two years before the 1990 Census. The two distributions, $p(y|\text{match})$ and $p(y|\text{non-match})$, are mostly distinct—meaning that in most cases it is possible to identify a candidate as a match or not given the score alone—but with some overlap.

In our application dataset, we do not know the match status. Thus we are faced with a single combined histogram from which we estimate the two component distributions and the proportion of the population of scores that belong to each component. Under the mixture model, the distribution of scores can be written as,

$$p(y) = \Pr(\text{match})\, p(y|\text{match}) + \Pr(\text{non-match})\, p(y|\text{non-match}). \qquad (1.6)$$

The mixture probability (Pr(match)) and the parameters of the distributions of matches ($p(y|\text{match})$) and non-matches ($p(y|\text{non-match})$) are estimated using the mixture model approach (as described in Chapter 18) applied to the combined histogram from the data with unknown match status.

To use the method to make record-linkage decisions, we construct a curve giving the false-match rate as a function of the decision threshold, the score

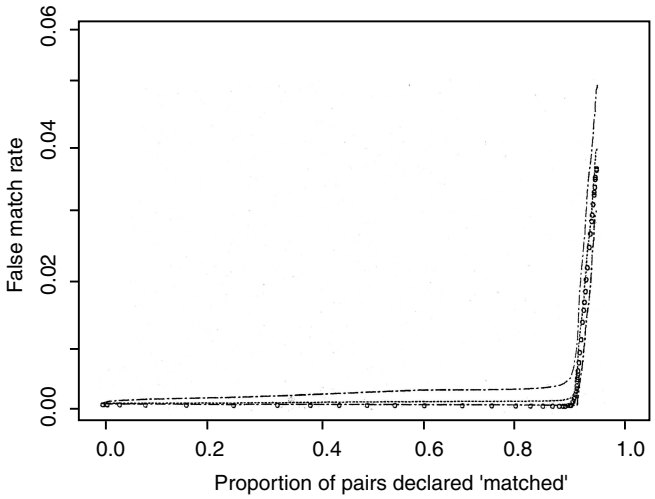

Figure 1.4 *Lines show expected false-match rate (and 95% bounds) as a function of the proportion of cases declared matches, based on the mixture model for record linkage. Dots show the actual false-match rate for the data.*

above which pairs will be 'declared' a match. For a given decision threshold, the probability distributions in (1.6) can be used to estimate the probability of a false match, a score y above the threshold originating from the distribution $p(y|\text{non-match})$. The lower the threshold, the more pairs we will declare as matches. As we declare more matches, the proportion of errors increases. The approach described here should provide an objective error estimate for each threshold. (See the validation in the next paragraph.) Then a decision maker can determine the threshold that provides an acceptable balance between the goals of declaring more matches automatically (thus reducing the clerical labor) and making fewer mistakes.

External validation of the probabilities using test data

The approach described above was externally validated using data for which the match status is known. The method was applied to data from three different locations of the 1988 test Census, and so three tests of the methods were possible. We provide detailed results for one; results for the other two were similar. The mixture model was fitted to the scores of all the candidate pairs at a test site. Then the estimated model was used to create the lines in Figure 1.4, which show the expected false-match rate (and uncertainty bounds) in terms of the proportion of cases declared matched, as the threshold varies from very high (thus allowing no matches) to very low (thus declaring almost all the candidate pairs to be matches). The false-match proportion is an increasing function of the number of declared matches, which makes sense: as

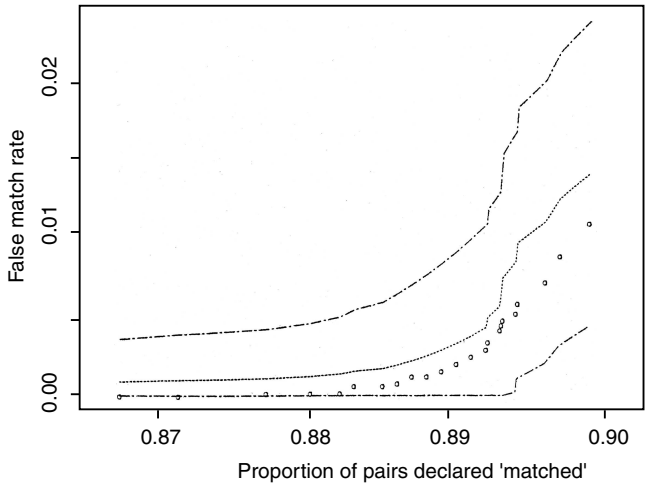

Figure 1.5 *Expansion of Figure 1.4 in the region where the estimated and actual match rates change rapidly. In this case, it would seem a good idea to match about 88% of the cases and send the rest to followup.*

we move rightward on the graph, we are declaring weaker and weaker cases to be matches.

The lines on Figure 1.4 display the expected proportion of false matches and 95% posterior bounds for the false-match rate as estimated from the model. (These bounds give the estimated range within which there is 95% posterior probability that the false-match rate lies. The concept of posterior intervals is discussed in more detail in the next chapter.) The dots in the graph display the actual false-match proportions, which track well with the model. In particular, the model would suggest a recommendation of declaring something less than 90% of cases as matched and giving up on the other 10% or so, so as to avoid most of the false matches, and the dots show a similar pattern.

It is clearly possible to match large proportions of the files with little or no error. Also, the quality of candidate matches becomes dramatically worse at some point where the false-match rate accelerates. Figure 1.5 takes a magnifying glass to the previous display to highlight the behavior of the calibration procedure in the region of interest where the false-match rate accelerates. The predicted false-match rate curves bend upward, close to the points where the observed false-match rate curves rise steeply, which is a particularly encouraging feature of the calibration method. The calibration procedure performs well from the standpoint of providing predicted probabilities that are close to the true probabilities and interval estimates that are informative and include the true values. By comparison, the original estimates of match probabilities, constructed by multiplying weights without empirical calibration, were highly inaccurate.

1.8 Some useful results from probability theory

We assume the reader is familiar with elementary manipulations involving probabilities and probability distributions. In particular, basic probability background that must be well understood for key parts of the book includes the manipulation of joint densities, the definition of simple moments, the transformation of variables, and methods of simulation. In this section we briefly review these assumed prerequisites and clarify some further notational conventions used in the remainder of the book. Appendix A provides information on some commonly used probability distributions.

As introduced in Section 1.3, we generally represent joint distributions by their joint probability mass or density function, with dummy arguments reflecting the name given to each variable being considered. Thus for two quantities u and v, we write the joint density as $p(u, v)$; if specific values need to be referenced, this notation will be further abused as with, for example, $p(u, v=1)$.

In Bayesian calculations relating to a joint density $p(u, v)$, we will often refer to a *conditional* distribution or density function such as $p(u|v)$ and a *marginal* density such as $p(u) = \int p(u, v)dv$. In this notation, either or both u and v can be vectors. Typically it will be clear from the context that the range of integration in the latter expression refers to the entire range of the variable being integrated out. It is also often useful to *factor* a joint density as a product of marginal and conditional densities; for example, $p(u, v, w) = p(u|v, w)p(v|w)p(w)$.

Some authors use different notations for distributions on parameters and observables—for example, $\pi(\theta), f(y|\theta)$—but this obscures the fact that all probability distributions have the same *logical* status in Bayesian inference. We must always be careful, though, to indicate appropriate conditioning; for example, $p(y|\theta)$ is different from $p(y)$. In the interests of conciseness, however, our notation hides the conditioning on hypotheses that hold throughout—no probability judgments can be made in a vacuum—and to be more explicit one might use a notation such as the following:

$$p(\theta, y|H) = p(\theta|H)p(y|\theta, H),$$

where H refers to the set of hypotheses or assumptions used to define the model. Also, we sometimes suppress explicit conditioning on known explanatory variables, x.

We use the standard notations, $\mathrm{E}(\cdot)$ and $\mathrm{var}(\cdot)$, for mean and variance, respectively:

$$\mathrm{E}(u) = \int up(u)du, \quad \mathrm{var}(u) = \int (u - \mathrm{E}(u))^2 p(u)du.$$

For a vector parameter u, the expression for the mean is the same, and the covariance matrix is defined as

$$\mathrm{var}(u) = \int (u - \mathrm{E}(u))(u - \mathrm{E}(u))^T p(u)du,$$

where u is considered a column vector. (We use the terms 'variance matrix' and 'covariance matrix' interchangeably.) This notation is slightly imprecise, because $\mathrm{E}(u)$ and $\mathrm{var}(u)$ are really functions of the distribution function, $p(u)$, not of the variable u. In an expression involving an expectation, any variable that does not appear explicitly as a conditioning variable is assumed to be integrated out in the expectation; for example, $\mathrm{E}(u|v)$ refers to the conditional expectation of u with v held fixed—that is, the conditional expectation as a function of v—whereas $\mathrm{E}(u)$ is the expectation of u, averaging over v (as well as u).

Modeling using conditional probability

Useful probability models often express the distribution of observables conditionally or hierarchically rather than through more complicated unconditional distributions. For example, suppose y is the height of a university student selected at random. The marginal distribution $p(y)$ is (essentially) a mixture of two approximately normal distributions centered around 160 and 175 centimeters. A more useful description of the distribution of y would be based on the joint distribution of height and sex: $p(\text{male}) \approx p(\text{female}) \approx \frac{1}{2}$, along with the conditional specifications that $p(y|\text{female})$ and $p(y|\text{male})$ are each approximately normal with means 160 and 175 cm, respectively. If the conditional variances are not too large, the marginal distribution of y is bimodal. In general, we prefer to model complexity with a hierarchical structure using additional variables rather than with complicated marginal distributions, even when the additional variables are unobserved or even unobservable; this theme underlies mixture models, as discussed in Chapter 18. We repeatedly return to the theme of conditional modeling throughout the book.

Means and variances of conditional distributions

It is often useful to express the mean and variance of a random variable u in terms of the conditional mean and variance given some related quantity v. The mean of u can be obtained by averaging the conditional mean over the marginal distribution of v,

$$\mathrm{E}(u) = \mathrm{E}(\mathrm{E}(u|v)), \tag{1.7}$$

where the inner expectation averages over u, conditional on v, and the outer expectation averages over v. Identity (1.7) is easy to derive by writing the expectation in terms of the joint distribution of u and v and then factoring the joint distribution:

$$\mathrm{E}(u) = \int\int up(u,v)dudv = \int\int u\, p(u|v)du\, p(v)dv = \int \mathrm{E}(u|v)p(v)dv.$$

The corresponding result for the variance includes two terms, the mean of the conditional variance and the variance of the conditional mean:

$$\text{var}(u) = \text{E}(\text{var}(u|v)) + \text{var}(\text{E}(u|v)). \tag{1.8}$$

This result can be derived by expanding the terms on the right side of (1.8):

$$\begin{aligned} &\text{E}[\text{var}(u|v)] + \text{var}[\text{E}(u|v)] \\ &\quad = \text{E}[\text{E}(u^2|v) - (\text{E}(u|v))^2] + \text{E}[(\text{E}(u|v))^2] - (\text{E}[\text{E}(u|v)])^2 \\ &\quad = \text{E}(u^2) - \text{E}[(\text{E}(u|v))^2] + \text{E}[(\text{E}(u|v))^2] - (\text{E}(u))^2 \\ &\quad = \text{E}(u^2) - (\text{E}(u))^2 = \text{var}(u). \end{aligned}$$

Identities (1.7) and (1.8) also hold if u is a vector, in which case $\text{E}(u)$ is a vector and $\text{var}(u)$ a matrix.

Transformation of variables

It is common to transform a probability distribution from one parameterization to another. We review the basic result here for a probability density on a transformed space. For clarity, we use subscripts here instead of our usual generic notation, $p(\cdot)$. Suppose $p_u(u)$ is the density of the vector u, and we transform to $v = f(u)$, where v has the same number of components as u.

If p_u is a discrete distribution, and f is a one-to-one function, then the density of v is given by

$$p_v(v) = p_u(f^{-1}(v)).$$

If f is a many-to-one function, then a sum of terms appears on the right side of this expression for $p_v(v)$, with one term corresponding to each of the branches of the inverse function.

If p_u is a continuous distribution, and $v = f(u)$ is a one-to-one transformation, then the joint density of the transformed vector is

$$p_v(v) = |J|\, p_u(f^{-1}(v))$$

where $|J|$ is the determinant of the Jacobian of the transformation $u = f^{-1}(v)$ as a function of v; the Jacobian J is the square matrix of partial derivatives (with dimension given by the number of components of u), with the (i, j)th entry equal to $\partial u_i / \partial v_j$. Once again, if f is many-to-one, then $p_v(v)$ is a sum or integral of terms.

In one dimension, we commonly use the logarithm to transform the parameter space from $(0, \infty)$ to $(-\infty, \infty)$. When working with parameters defined on the open unit interval, $(0, 1)$, we often use the logistic transformation:

$$\text{logit}(u) = \log\left(\frac{u}{1-u}\right), \tag{1.9}$$

whose inverse transformation is

$$\text{logit}^{-1}(v) = \frac{e^v}{1+e^v}.$$

Another common choice is the probit transformation, $\Phi^{-1}(u)$, where Φ is the standard normal cumulative distribution function, to transform from $(0, 1)$ to $(-\infty, \infty)$.

1.9 Summarizing inferences by simulation

Simulation forms a central part of much applied Bayesian analysis, because of the relative ease with which samples can often be generated from a probability distribution, even when the density function cannot be explicitly integrated. In performing simulations, it is helpful to consider the duality between a probability density function and a histogram of a set of random draws from the distribution: given a large enough sample, the histogram can provide practically complete information about the density, and in particular, various sample moments, percentiles, and other summary statistics provide estimates of any aspect of the distribution, to a level of precision that can be estimated. For example, to estimate the 95th percentile of the distribution of θ, draw a random sample of size L from $p(\theta)$ and use the $0.95L$th order statistic. For most purposes, $L = 1000$ is adequate for estimating the 95th percentile in this way.

Another advantage of simulation is that extremely large or small simulated values often flag a problem with model specification or parameterization (for example, see Figure 4.2) that might not be noticed if estimates and probability statements were obtained in analytic form.

Generating values from a probability distribution is often straightforward with modern computing techniques based on (pseudo)random number sequences. A well-designed pseudorandom number generator yields a deterministic sequence that appears to have the same properties as a sequence of independent random draws from the uniform distribution on $[0, 1]$. Appendix A describes methods for drawing random samples from some commonly used distributions.

Sampling using the inverse cumulative distribution function.

As an introduction to the ideas of simulation, we describe a method for sampling from discrete and continuous distributions using the inverse cumulative distribution function. The *cumulative distribution function*, or *cdf*, F, of a one-dimensional distribution, $p(v)$, is defined by

$$\begin{aligned} F(v_*) &= \Pr(v \le v_*) \\ &= \begin{cases} \sum_{v \le v_*} p(v) & \text{if } p \text{ is discrete} \\ \int_{-\infty}^{v_*} p(v)dv & \text{if } p \text{ is continuous.} \end{cases} \end{aligned}$$

The inverse cdf can be used to obtain random samples from the distribution p, as follows. First draw a random value, U, from the uniform distribution on $[0, 1]$, using a table of random numbers or, more likely, a random number function on the computer. Now let $v = F^{-1}(U)$. The function F is not necessarily one-to-one—certainly not if the distribution is discrete—but $F^{-1}(U)$

Simulation draw	Parameters			Predictive quantities		
	θ_1	$\dots$	θ_k	$\tilde{y}_1$	$\dots$	$\tilde{y}_n$
1	θ_1^1	$\dots$	θ_k^1	$\tilde{y}_1^1$	$\dots$	$\tilde{y}_n^1$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
L	θ_1^L	$\dots$	θ_k^L	$\tilde{y}_1^L$	$\dots$	$\tilde{y}_n^L$

Table 1.1 *Structure of posterior and posterior predictive simulations. The superscripts are indexes, not powers.*

is unique with probability 1. The value v will be a random draw from p, and is easy to compute as long as $F^{-1}(U)$ is simple. For a discrete distribution, F^{-1} can simply be tabulated.

For a continuous example, suppose v has an exponential distribution with parameter λ (see Appendix A); then its cdf is $F(v) = 1 - \exp(-\lambda v)$, and the value of v for which $U = F(v)$ is $v = -\log(1-U)/\lambda$. Of course $1 - U$ also has the uniform distribution on $[0, 1]$, so we can obtain random draws from the exponential distribution as $-(\log U)/\lambda$. We discuss other methods of simulation in Part III of the book and Appendix A.

Simulation of posterior and posterior predictive quantities

In practice, we are most often interested in simulating draws from the posterior distribution of the model parameters θ, and perhaps from the posterior predictive distribution of unknown observables $\tilde{y}$. Results from a set of L simulation draws can be stored in the computer in an array, as illustrated in Table 1.1. We use the notation $l = 1, \dots, L$ to index simulation draws; $(\theta^l, \tilde{y}^l)$ is the corresponding joint draw of parameters and predicted quantities from their joint posterior distribution.

From these simulated values, we can estimate the posterior distribution of any quantity of interest, such as θ_1/θ_3, by just computing a new column in Table 1.1 using the existing L draws of $(\theta, \tilde{y})$. We can estimate the posterior probability of any event, such as $\Pr(\tilde{y}_1 + \tilde{y}_2 > \exp(\theta_1))$, by the proportion of the L simulations for which it is true. We are often interested in posterior intervals; for example, the central 95% posterior interval $[a, b]$ for the parameter θ_j, for which $\Pr(\theta_j < a) = 0.025$ and $\Pr(\theta_j > b) = 0.025$. These values can be directly estimated by the appropriate simulated values of θ_j, for example, the 25th and 976th order statistics if $L = 1000$. We commonly summarize inferences by 50% and 95% intervals.

We return to the accuracy of simulation inferences in Section 10.2 after we have gained some experience using simulations of posterior distributions in some simple examples.

1.10 Computation and software

At the time of writing, the authors rely primarily on the statistical package `R` for graphs and basic simulations, fitting of classical simple models (including regression, generalized linear models, and nonparametric methods such as locally-weighted regression), optimization, and some simple programming.

We use the Bayesian inference package `Bugs` (using WinBugs run directly from within `R`; see Appendix C) as a first try for fitting most models. If there are difficulties in setting up the model or achieving convergence of the simulations, we explore the model more carefully in `R` and, if necessary for computational speed, a lower-level language such as Fortran or C (both of which can be linked from `R`). In any case, we typically work within `R` to plot and transform the data before model fitting, and to display inferences and model checks afterwards.

Specific computational tasks that arise in Bayesian data analysis include:

- Vector and matrix manipulations (see Table 1.1)
- Computing probability density functions (see Appendix A)
- Drawing simulations from probability distributions (see Appendix A for standard distributions and Exercise 1.9 for an example of a simple stochastic process)
- Structured programming (including looping and customized functions)
- Calculating the linear regression estimate and variance matrix (see Chapter 14)
- Graphics, including scatterplots with overlain lines and multiple graphs per page (see Chapter 6 for examples).

Our general approach to computation is to fit many models, gradually increasing the complexity. We do *not* recommend the strategy of writing a model and then letting the computer run overnight to estimate it perfectly. Rather, we prefer to fit each model relatively quickly, using inferences from the previously-fitted simpler models as starting values, and displaying inferences and comparing to data before continuing.

We discuss computation in detail in Part III of this book after first introducing the fundamental concepts of Bayesian modeling, inference, and model checking. Appendix C illustrates how to perform computations in `R` and `Bugs` in several different ways for a single example.

1.11 Bibliographic note

Several good introductory books have been written on Bayesian statistics, beginning with Lindley (1965). Berry (1996) presents, from a Bayesian perspective, many of the standard topics for an introductory statistics textbook. Congdon (2001, 2003) and Gill (2002) are recent introductory books on applied Bayesian statistics that use the statistical package `Bugs`. Carlin and

Louis (2001) cover the theory and applications of Bayesian inference, focusing on biological applications and connections to classical methods.

The bibliographic notes at the ends of the chapters in this book refer to a variety of specific applications of Bayesian data analysis. Several review articles in the statistical literature, such as Breslow (1990) and Racine et al. (1986), have appeared that discuss, in general terms, areas of application in which Bayesian methods have been useful. The volumes edited by Gatsonis et al. (1993–2002) are collections of Bayesian analyses, including extensive discussions about choices in the modeling process and the relations between the statistical methods and the applications.

The foundations of probability and Bayesian statistics are an important topic that we treat only very briefly. Bernardo and Smith (1994) give a thorough review of the foundations of Bayesian models and inference with a comprehensive list of references. Jeffreys (1961) is a self-contained book about Bayesian statistics that comprehensively presents an inductive view of inference; Good (1950) is another important early work. Jaynes (1983) is a collection of reprinted articles that present a deductive view of Bayesian inference, which we believe is quite similar to ours. Both Jeffreys and Jaynes focus on applications in the physical sciences. Jaynes (1996) focuses on connections between statistical inference and the philosophy of science and includes several examples of physical probability.

De Finetti (1974) is an influential work that focuses on the crucial role of exchangeability. More approachable discussions of the role of exchangeability in Bayesian inference are provided by Lindley and Novick (1981) and Rubin (1978a, 1987a). The non-Bayesian article by Draper et al. (1993) makes an interesting attempt to explain how exchangeable probability models can be justified in data analysis. Berger and Wolpert (1984) give a comprehensive discussion and review of the likelihood principle, and Berger (1985, Sections 1.6, 4.1, and 4.12) reviews a range of philosophical issues from the perspective of Bayesian decision theory.

Pratt (1965) and Rubin (1984) discuss the relevance of Bayesian methods for applied statistics and make many connections between Bayesian and non-Bayesian approaches to inference. Further references on the foundations of statistical inference appear in Shafer (1982) and the accompanying discussion. Kahneman, Slovic, and Tversky (1982) present the results of various psychological experiments that assess the meaning of 'subjective probability' as measured by people's stated beliefs and observed actions. Lindley (1971a) surveys many different statistical ideas, all from the Bayesian perspective. Box and Tiao (1973) is an early book on applied Bayesian methods. They give an extensive treatment of inference based on normal distributions, and their first chapter, a broad introduction to Bayesian inference, provides a good counterpart to Chapters 1 and 2 of this book.

The iterative process involving modeling, inference, and model checking that we present in Section 1.1 is discussed at length in the first chapter of Box

and Tiao (1973) and also in Box (1980). Cox and Snell (1981) provide a more introductory treatment of these ideas from a less model-based perspective.

Many good books on the mathematical aspects of probability theory are available, such as Feller (1968) and Ross (1983); these are useful when constructing probability models and working with them. O'Hagan (1988) has written an interesting introductory text on probability from an explicitly Bayesian point of view.

Physical probability models for coin tossing are discussed by Keller (1986), Jaynes (1996), and Gelman and Nolan (2002b). The football example of Section 1.6 is discussed in more detail in Stern (1991); see also Harville (1980) and Glickman (1993) and Glickman and Stern (1998) for analyses of football scores not using the point spread. Related analyses of sports scores and betting odds appear in Stern (1997, 1998). For more background on sports betting, see Snyder (1975) and Rombola (1984).

An interesting real-world example of probability assignment arose with the explosion of the Challenger space shuttle in 1986; Martz and Zimmer (1992), Dalal, Fowlkes, and Hoadley (1989), and Lavine (1991) present and compare various methods for assigning probabilities for space shuttle failures. (At the time of writing we are not aware of similar contributions relating to the latest space shuttle tragedy in 2003.) The record-linkage example in Section 1.7 appears in Belin and Rubin (1995b), who discuss the mixture models and calibration techniques in more detail. The Census problem that motivated the record linkage is described by Hogan (1992).

In all our examples, probabilities are assigned using statistical modeling and estimation, not by 'subjective' assessment. Dawid (1986) provides a general discussion of probability assignment, and Dawid (1982) discusses the connections between calibration and Bayesian probability assignment.

The graphical method of jittering, used in Figures 1.1 and 1.2 and elsewhere in this book, is discussed in Chambers et al. (1983). For information on the statistical packages `R` and `BUGS`, see Becker, Chambers, and Wilks (1988), R Project (2002), Fox (2002), Venables and Ripley (2002), and Spiegelhalter et al. (1994, 2003).

1.12 Exercises

1. Conditional probability: suppose that if $\theta = 1$, then y has a normal distribution with mean 1 and standard deviation σ, and if $\theta = 2$, then y has a normal distribution with mean 2 and standard deviation σ. Also, suppose $\Pr(\theta = 1) = 0.5$ and $\Pr(\theta = 2) = 0.5$.

   (a) For $\sigma = 2$, write the formula for the marginal probability density for y and sketch it.

   (b) What is $\Pr(\theta = 1|y = 1)$, again supposing $\sigma = 2$?

   (c) Describe how the posterior density of θ changes in shape as σ is increased and as it is decreased.

2. Conditional means and variances: show that (1.7) and (1.8) hold if u is a vector.

3. Probability calculation for genetics (from Lindley, 1965): suppose that in each individual of a large population there is a pair of genes, each of which can be either x or X, that controls eye color: those with xx have blue eyes, while heterozygotes (those with Xx or xX) and those with XX have brown eyes. The proportion of blue-eyed individuals is p^2 and of heterozygotes is $2p(1-p)$, where $0 < p < 1$. Each parent transmits one of its own genes to the child; if a parent is a heterozygote, the probability that it transmits the gene of type X is $\frac{1}{2}$. Assuming random mating, show that among brown-eyed children of brown-eyed parents, the expected proportion of heterozygotes is $2p/(1+2p)$. Suppose Judy, a brown-eyed child of brown-eyed parents, marries a heterozygote, and they have n children, all brown-eyed. Find the posterior probability that Judy is a heterozygote and the probability that her first grandchild has blue eyes.

4. Probability assignment: we will use the football dataset to estimate some conditional probabilities about professional football games. There were twelve games with point spreads of 8 points; the outcomes in those games were: $-7, -5, -3, -3, 1, 6, 7, 13, 15, 16, 20, 21$, with positive values indicating wins by the favorite and negative values indicating wins by the underdog. Consider the following conditional probabilities:

   $\Pr(\text{favorite wins} \,|\, \text{point spread} = 8)$,

   $\Pr(\text{favorite wins by at least } 8 \,|\, \text{point spread} = 8)$,

   $\Pr(\text{favorite wins by at least } 8 \,|\, \text{point spread} = 8 \text{ and favorite wins})$.

   (a) Estimate each of these using the relative frequencies of games with a point spread of 8.

   (b) Estimate each using the normal approximation for the distribution of (outcome − point spread).

5. Probability assignment: the 435 U.S. Congress members are elected to two-year terms; the number of voters in an individual Congressional election varies from about 50,000 to 350,000. We will use various sources of information to estimate roughly the probability that at least one Congressional election is tied in the next national election.

   (a) Use any knowledge you have about U.S. politics. Specify clearly what information you are using to construct this conditional probability, even if your answer is just a guess.

   (b) Use the following information: in the period 1900–1992, there were 20,597 Congressional elections, out of which 6 were decided by fewer than 10 votes and 49 decided by fewer than 100 votes.

   See Gelman, King, and Boscardin (1998), Mulligan and Hunter (2001), and Gelman, Katz, and Tuerlinckx (2002) for more on this topic.

6. Conditional probability: approximately 1/125 of all births are fraternal twins and 1/300 of births are identical twins. Elvis Presley had a twin brother (who died at birth). What is the probability that Elvis was an identical twin? (You may approximate the probability of a boy or girl birth as $\frac{1}{2}$.)

7. Conditional probability: the following problem is loosely based on the television game show *Let's Make a Deal*. At the end of the show, a contestant is asked to choose one of three large boxes, where one box contains a fabulous prize and the other two boxes contain lesser prizes. After the contestant chooses a box, Monty Hall, the host of the show, opens one of the two boxes containing smaller prizes. (In order to keep the conclusion suspenseful, Monty does not open the box selected by the contestant.) Monty offers the contestant the opportunity to switch from the chosen box to the remaining unopened box. Should the contestant switch or stay with the original choice? Calculate the probability that the contestant wins under each strategy. This is an exercise in being clear about the information that should be conditioned on when constructing a probability judgment. See Selvin (1975) and Morgan et al. (1991) for further discussion of this problem.

8. Subjective probability: discuss the following statement. 'The probability of event E is considered "subjective" if two rational persons A and B can assign unequal probabilities to E, $P_A(E)$ and $P_B(E)$. These probabilities can also be interpreted as "conditional": $P_A(E) = P(E|I_A)$ and $P_B(E) = P(E|I_B)$, where I_A and I_B represent the knowledge available to persons A and B, respectively.' Apply this idea to the following examples.

   (a) The probability that a '6' appears when a fair die is rolled, where A observes the outcome of the die roll and B does not.

   (b) The probability that Brazil wins the next World Cup, where A is ignorant of soccer and B is a knowledgeable sports fan.

9. Simulation of a queuing problem: a clinic has three doctors. Patients come into the clinic at random, starting at 9 a.m., according to a Poisson process with time parameter 10 minutes: that is, the time after opening at which the first patient appears follows an exponential distribution with expectation 10 minutes and then, after each patient arrives, the waiting time until the next patient is independently exponentially distributed, also with expectation 10 minutes. When a patient arrives, he or she waits until a doctor is available. The amount of time spent by each doctor with each patient is a random variable, uniformly distributed between 5 and 20 minutes. The office stops admitting new patients at 4 p.m. and closes when the last patient is through with the doctor.

   (a) Simulate this process once. How many patients came to the office? How many had to wait for a doctor? What was their average wait? When did the office close?

(b) Simulate the process 100 times and estimate the median and 50% interval for each of the summaries in (a).