4 *Data Analysis and Uncertainty*

4.1 Introduction

In this chapter, we focus on uncertainty and how to cope with it. Not only is the process of mapping from the real world to our databases seldom perfect, but the domain of the mapping—the real world itself—is beset with ambiguities and uncertainties. The basic tool for dealing with uncertainty is probability, and we begin by defining the concept and showing how it is used to construct statistical models. Section 4.2 provides a brief discussion of the distinction between probability calculus and the interpretation of probability, focusing on the two main interpretations: the frequentist and the subjective (Bayesian). Section 4.3 extends this discussion to define the concept of a random variable, with a particular focus on the relationships that can exist between multiple random variables.

Fundamental to many data mining activities is the notion of a sample. Sometimes the database contains only a sample from the universe of possible records; section 4.4 explores this situation, explaining why samples are often sufficient to work with. Section 4.5 describes *estimation*, the process of moving beyond a data sample to develop parameter estimates for a model describing the data. In particular, we review in some detail the basic principles of the maximum likelihood and Bayesian approaches to estimation. Section 4.6 discusses the closely related topic of how to evaluate the quality of a hypothesis on the basis of observed data. Section 4.7 outlines various systematic methods for drawing samples from data. Section 4.8 presents some concluding remarks, and section 4.9 gives pointers to more detailed reading.

4.2 Dealing with Uncertainty

The ubiquity of the idea of uncertainty is illustrated by the rich variety of words used to describe it and related concepts. *Probability, chance, randomness, luck, hazard,* and *fate* are just a few examples. The omnipresence of uncertainty requires us to be able to cope with it: modeling uncertainty is a necessary component of almost all data analysis. Indeed, in some cases our primary aim is to model the uncertain or random aspects of data. It is one of the great achievements of science that we have developed a deep and powerful understanding of uncertainty. The capricious gods that were previously invoked to explain the lack of predictability in the world have been replaced by mathematical, statistical, and computer-based models that allow us to understand and manipulate uncertain events. We can even attempt the seemingly impossible and predict uncertain events, where prediction for a data miner either can mean the prediction of future events (where the notion of uncertainty is very familiar) or prediction in a nontemporal sense of a variable whose true value is somehow hidden from us (for example, diagnosing whether a person has cancer, based on only descriptive symptoms).

We may be uncertain for various reasons. Our data may be only a sample from the population we wish to study, so that we are uncertain about the extent to which different samples differ from each other and from the overall population. Perhaps our interest lies in making a prediction about tomorrow, based on the data we have today, so that our conclusions are subject to uncertainty about what the future will bring. Perhaps we are ignorant and cannot observe some value, and have to base our ideas on our "best guess" about it. And so on.

Many conceptual bases have been formulated for handling uncertainty and ignorance. Of these, by far the most widely used is probability. Fuzzy logic is another that has a moderately large following, but this area—along with closely related areas such as possibility theory and rough sets—remains rather controversial: it lacks the sound theoretical backbone and widespread application and acceptance of probability. These ideas may one day develop solid foundations, and become widely used, but because of their current uncertain status we will not consider them further in this book.

It is useful to distinguish between *probability theory* and *probability calculus*. The former is concerned with the interpretation of probability while the latter is concerned with the manipulation of the mathematical representation of probability. (Unfortunately, not all textbooks make this distinction between the two terms—often books on probability calculus are given titles such as

"Introduction to the Theory of Probability.") The distinction is an important one because it permits the separation of those areas about which there is universal agreement (the calculus) from those areas about which opinions differ (the theory). The calculus is a branch of mathematics, based on well-defined and generally accepted axioms (stated by the Russian mathematician Kolmogorov in the 1930s); the aim is to explore the consequences of those axioms. (There are some areas in which different sets of axioms are used, but these are rather specialized and generally do not impinge on problems of data mining.) The theory, on the other hand, leaves scope for perspectives on the mapping from the real world to the mathematical representation—i.e., on what probability is.

A study of the history and philosophy of probability theory reveals that there are as many perspectives on the meaning of probability as there are thinkers. However, the views can be grouped into variants of a few different types. Here we shall restrict ourselves to discussing the two most important types (in terms of their impact on data mining practice). More philosophically inclined readers may wish to consult section 4.9 for references to material containing broader discussions.

The *frequentist view* of probability takes the perspective that probability is an objective concept. In particular, the probability of an event is defined as the limiting proportion of times that the event would occur in repetitions of essentially identical situations. A simple example is the proportion of times a head comes up in repeatedly tossing a coin. This interpretation restricts our application of probability: for instance we cannot assess the probability that a particular athlete will win a medal in the next Olympics because this is a one-off event, where the notion of a "limiting proportion" makes no sense. On the other hand, we can certainly assess the probability that a customer in a supermarket will purchase a certain item, since we can use a large number of similar customers as the basis for a limiting proportion argument. It is clear in this last example that some *idealization* is going on: different customers are not really the same as repetitions of a single customer. As in all scientific modeling we need to decide what aspects are important for our model to be sufficiently accurate. In predicting customer behavior we might decide that the differences between customers do not matter.

The frequentist view was the dominant perspective on probability throughout most of the last century, and hence it underpins most widely used statistical software. However, in the last decade or so, a competing view has acquired increasing importance. This view, that of *subjective probability*, has been around since people first started formalizing probabilistic notions, but

until recently it was primarily of theoretical interest. What revived the approach was the development of the computer and of powerful algorithms for manipulating and processing subjective probabilities. The principles and methodologies for data analysis that derive from the subjective point of view are often referred to as *Bayesian statistics*. A central tenet of Bayesian statistics is the explicit characterization of *all* forms of uncertainty in a data analysis problem, including uncertainty about any parameters we estimate from the data, uncertainty as to which among a set of model structures are best or closest to "truth," uncertainty in any forecast we might make, and so on. Subjective probability is a very flexible framework for modeling such uncertainty in different forms.

From the perspective of subjective probability, probability is an individual degree of belief that a given event will occur. Thus, probability is not an objective property of the outside world, but rather an internal state of the individual—and may differ from individual to individual. Fortunately it turns out that if we adopt certain tenets of rational behaviour the set of axioms underlying subjective probability is the same as that underlying the frequentist view. The *calculus* is the same for the two viewpoints, even though the underlying *interpretation* is quite different.

Of course, this does not imply that the conclusions drawn using the two approaches are necessarily the same. At the very least, subjective probability can make statements about areas that frequentist probability cannot address. Moreover, statistical inferences based on subjective probability necessarily involve a subjective component—the initial or prior belief that an event will happen. As noted above, this factor is likely to differ from person to person.

Nonetheless, the frequentist and subjective viewpoints in many cases lead to roughly the same answers, particularly for simple hypotheses and large data sets. Rather than committing to one viewpoint or the other, many practitioners view both as useful in their own right, with each appropriate in different situations. The methodologies for data analysis that derive from the frequentist view tend to be computationally simpler, and thus (to date at least) have dominated in the development of data mining techniques where the size of the data sets do not favor the application of complex computational methods. However, when applied with care the Bayesian (subjective) methodology has the ability to tease out more subtle information from the data. Just as applied statistics has seen increased interest in Bayesian methods in recent years, we can expect to see more Bayesian ideas being applied in data mining in the future. In the rest of this book we will refer to both frequentist and Bayesian views where appropriate. As we will see later in this

chapter, in a certain sense the two viewpoints can be reconciled: the frequentist methodology of fitting models and patterns to data can be implemented as a special case of a more general Bayesian methodology. For the practitioner this is quite useful, since it means that the same general modeling and computational apparatus can be used.

4.3 Random Variables and Their Relationships

We introduced the notion of a variable in chapter 2. In this chapter we introduce the concept of a *random variable*. A random variable is a mapping from a property of objects to a variable that can take one of a set of possible values, via a process that appears to the observer to have some element of unpredictability to it. The possible values of a random variable X are called the domain of X. We use uppercase letters such as X to refer to a random variable and lowercase letters such as x to refer to a value of a random variable.

An example of a random variable is the outcome of a coin toss (the domain is the set {heads, tails}). Less obvious examples of random variables include the number of times we have to toss a coin to obtain the first head (the domain is the set of positive integers) and the flying time of a paper aeroplane in seconds (the domain is the set of positive real numbers).

The appendix defines the basic properties of univariate (single) random variables, including both probability mass functions $p(X)$ when the domain of X is finite and probability density functions $f(x)$ when the domain of X is the real-line or any interval defined on it. Basic properties of the expectation of X, $E[X] = \int xf(x)dx$, for real-valued X, are also reviewed, noting for example that since E is a linear operator we have that $E[X+Y] = E[X]+E[Y]$. These basic properties are extremely useful in allowing us to derive general principles for data analysis in a statistical context and we will refer to distributions, densities, expectation, etc., frequently throughout the remainder of this chapter.

4.3.1 Multivariate Random Variables

Since data mining often deals with multiple variables, we must also introduce the concept of a *multivariate random variable*. A multivariate random variable $\mathbf{X}$ is a set $X_1, \ldots, X_p$ of random variables. We use the m-dimensional vector $\mathbf{x} = \{x_1, \ldots, x_p\}$ to denote a set of values for $\mathbf{X}$. The *density function* $f(\mathbf{X})$ of the multivariate random variable $\mathbf{X}$ is called the *joint density func-*

tion of $\mathbf{X}$. We denote this as $f(\mathbf{X}) = f(X_1 = x_1, \ldots, X_p = x_p)$, or simply $f(x_1, \ldots, x_p)$. Similarly, we have joint probability distributions for variables taking values in a finite set. Note that $f(\mathbf{X})$ is a scalar function of p variables.

The density function of any single variable in the set $\mathbf{X}$ (or, more generally, any subset of the complete set of variables) is called a *marginal density* of the joint density. Technically, it is derived from the joint density by summing or integrating across the variables not included in the subset. For example, for a tri-variate random variable $\mathbf{X} = (X_1, X_2, X_3)$ the marginal density of $f(X_1)$ is given by $f(x_1) = \int \int f(x_1, x_2, x_3) dx_2 dx_3$.

The density of a single variable (or a subset of the complete set of variables) given (or "conditional on") particular values of the other variables is a *conditional density*. Thus we can speak of the conditional density of variable X_1 given that X_2 takes the value 6, denoted $f(x_1 \mid x_2 = 6)$. In general, the conditional density of X_1 given some value of X_2 is denoted by $f(x_1 \mid x_2)$, and is defined as

$$f(x_1 \mid x_2) = \frac{f(x_1, x_2)}{f(x_2)}. \tag{4.1}$$

For discrete-valued random variables we have equivalent definitions ($p(a_1 \mid a_2)$, etc.). We can also use mixtures of the two—e.g., a conditional probability density function $f(x_1 \mid a_1)$ for a continuous variable conditioned on a categorical variable, and a conditional probability mass function $p(a_1 \mid x_1)$ for the reverse case.

Example 4.1 Suppose we have data on purchases of products from supermarkets, with each observation (row) in the data matrix representing the products bought by one customer. Let each column represent a particular product, and associate a random variable with each column so that there is one variable per product. An observation in a given row and column has value 1 if the customer corresponding to that row bought the product from that column, and has value 0 otherwise.

Denote by A the binary random variable for a particular column, corresponding to the event "purchase of product A." A data-driven estimate of the probability that A takes value 1 is simply the fraction of customers who bought product A—i.e., n_A/n, where n is the total number of customers and n_A is the number of customers who bought product A. For example, if $n = 100,000$ and $n_A = 10,000$, an estimate of the probability that a randomly selected customer bought product A is 0.1.

Now consider a second product (a second column in the data matrix), with random variable B defined in the same way as A. Let n_B be the number of customers who bought product B; assume n_B = 5000 and therfore $p(B = 1) =$

0.05. Now let n_{AB} be the number of customers who purchased *both* A and B. Following the same argument as above, an estimate of $p(A = 1, B = 1)$ is given by n_{AB}/n. We can now estimate $p(B = 1|A = 1)$ as n_{AB}/n_A. Thus, for example, if $n_{AB} = 10$, we estimate $p(B = 1|A = 1)$ as $10/10,000 = 0.001$. We see from this that, while the estimated probability of a customer buying product B is 0.05, this reduces to 0.001 if we know that this customer bought product A as well. For the people in our database, the proportion of people buying B is far smaller among those who also bought A than among the people in the database as a whole (and thus smaller than among those who did not buy A). This prompts the question of whether buying A makes the purchase of B less likely in general, or whether this finding is simply an accident true only of the data we happen to have in our database. This is precisely the sort of question that we will address in the remainder of this chapter, particularly in section 4.6 on *hypothesis testing*.

Note that particular variables in the multivariate set $\mathbf{X}$ may well be related to each other in some manner. Indeed, a generic problem in data mining is to find relationships between variables. Is purchasing item A likely to be related to purchasing item B? Is detection of pattern A in the trace of a measuring instrument likely to be followed shortly afterward by a particular fault? Variables are said to be *independent* if there is no relationship between the occurrence of values of the variables; otherwise they are *dependent*. More formally, variables X and Y are independent if and only if $p(x, y) = p(x)p(y)$ for all values of X and Y. An equivalent formulation is that X and Y are independent if and only if $p(x \mid y) = p(x)$ or $p(y \mid x) = p(y)$ for all values of X and Y. (Note that these definitions hold whether each p in the expression is a probability mass function or a density function—in the latter case the variables are independent if and only if $f(x, y) = f(x)f(y)$). The second form of the definition shows that when X and Y are independent the distribution of X is the same whether or not the value of Y is known. Thus, Y carries no information about X, in the sense that the value taken by Y does not influence the probability of X taking any value. The random variables A and B in example 4.3.1 describing supermarket purchases are likely to be dependent, given the data as stated.

We can generalize these ideas to more than two variables. For example, we say that X is *conditionally independent* of Y given Z if for all values of X, Y, and Z we have that $p(x, y \mid z) = p(x \mid z)p(y \mid z)$, or equivalently $p(x \mid y, z) = p(x \mid z)$. To illustrate, suppose a person purchases bread (so that a random variable Z takes the value 1). Then subsequent purchases of butter (random variable X takes the value 1) and cheese (random variable

Y takes the value 1) might be modeled as being conditionally independent—the probability of purchasing cheese is unaffected by whether or not butter was purchased, once we know that bread has been purchased.

Note that conditional independence need not imply marginal (unconditional) independence. That is, the conditional independence relations above do not imply $p(x, y) = p(x)p(y)$. For example, in our illustration we might reasonably expect purchases of butter and cheese to be dependent in general (since they are both dependent on bread purchases). The reverse also applies: X and Y may be (unconditionally) independent, but conditionally dependent given a third variable Z. The subtleties of these dependence and independence relations have important consequences for data miners. In particular, even though two observed variables (such as butter and cheese) may appear to be dependent given the data, their true relationship may be masked by a third (potentially unobserved) variable (such as bread in our illustration).

Example 4.2 Care is needed when studying and interpreting conditional independence statements. Consider the following hypothetical example. A and B represent two different treatments, and the fractions shown in the table are the fraction of patients who recover (thus, at the top left, 2 out of 10 "old" patients receiving treatment A recover). The data have been partitioned into "old" and "young" groups, according to whether the patients were older or younger than 30.

	A	B
Old	2/10	30/90
Young	48/90	10/10

For each of the two age strata, treatment B appears superior to treatment A. However, now consider the overall results—obtained by aggregating the rows of the above table:

	A	B
Total	50/100	40/100

Overall, in this aggregate table, treatment A seems superior to treatment B. At first glance this result seems rather mysterious (in fact, it is known as *Simpson's paradox* (Simpson, 1951)).

The apparent contradiction between the two sets of results is explained by the fact that the first set is conditional on particular age strata, while the second is unconditional. When the two conditional statements are combined, the differences in sample sizes of the four groups cause the proportions based on the larger samples (Old B and Young A) to dominate the other two proportions.

The assumption of conditional independence is widely used in the context of sequential data, for which the next value in the sequence is often independent of all of the past values in the sequence given only the current value in the sequence. In this context, conditional independence is known as the *first-order Markov* property.

The notions of independence and conditional independence (which can be viewed as a generalization of independence) are central to many of the key concepts in data analysis, as we shall see in later chapters. The assumptions of independence and conditional independence enable us to factor the joint densities of many variables into much more tractable products of simpler densities, e.g.,

$$f(x_1, \ldots, x_n) = f(x_1) \prod_{j=2}^{n} f(x_j | x_{j-1}), \tag{4.2}$$

where each variable x_j is conditionally independent of variables $x_1, \ldots, x_{j-2}$, given the value of x_j (this is an example of a first-order Markov model). In addition to the computational benefits provided by such simplifications, it also provides important modeling gains by allowing us to construct more understandable models with fewer parameters. Nonetheless, independence is a very strong assumption that is frequently violated in practice (for example, assuming sequences of letters in text are first-order Markov may not be realistic). Still, keeping in mind that our models are inevitably approximations to the real world, the benefits of appropriate independence assumptions often outweigh the alternative of building more complex but less stable models. We will return to this theme of modeling in chapter 6.

A special case of dependency is correlation, or linear dependency, as introduced in chapter 2. (Note that statistical dependence is not the same as correlation: two variables may be dependent but not linearly correlated). Variables are said to be positively correlated if high values of one variable tend to be associated with high values of the other, and to be negatively correlated if high values of one tend to be associated with low values of the other. It is important not to confuse correlation with *causation*. Two variables may be highly positively correlated without any causal relationship between them. For example, yellow-stained fingers and lung cancer may be corre-

lated, but are causally linked only via a third variable, namely whether a person smokes or not. Similarly, human reaction time and earned income may be negatively correlated, but this does not mean that one causes the other. In this case a more convincing explanation is that a third variable, age, is causally related to both of these variables.

Example 4.3 A paper published in the *Journal of the American Medical Association* in 1987 (volume 257, page 785) examined the in-hospital mortality for 18,986 coronary bypass graft operations that were carried out at 77 hospitals in the United States. A regression analysis (see chapter 11) showed that hospitals that carried out more operations tended to have lower in-hospital mortality rates (even adjusting for different types of cases at different hospitals). From this pattern it was concluded that average in-hospital mortality following this type of operation would be reduced if the low-volume surgery units were closed.

However, determining the relationship between quality of outcome and number of treated cases in a hospital requires a longitudinal analysis in which the sizes are deliberately manipulated. The results of large-volume hospitals might degrade if their volume was increased. The correlation between outcome and size might have arisen not because larger size induces superior performance, but because superior performance attracts more cases, or because both the number of cases and the outcome are related to some other factor.

4.4 Samples and Statistical Inference

As we noted in chapter 2, many data mining problems involve the entire population of interest, while others involve just a sample from this population. In the latter case, the samples may arise at the start—perhaps only a sample of tax-payers is selected for detailed investigation; perhaps a complete census of the population is carried out only occasionally, with just a sample being selected in most years; or perhaps the data set consists of market research results. In other cases, even though the complete data set is available, the data mining operation is carried out on a sample. This is entirely legitimate if the aim is *modeling* (see chapter 1), which seeks to represent the prominent structures of the data, and not small idiosyncratic deviations. Such structures will be preserved in a sample, provided it is not too small. However, working with a small sample of a large data set may be less appropriate if the aim is *pattern detection*: in this case the aim may be to detect small deviations from the bulk of the data, and if the sample is too small such deviations may be excluded. Moreover, if the aim is to detect records that show anomalous behavior, the analysis must be based on the entire sample.

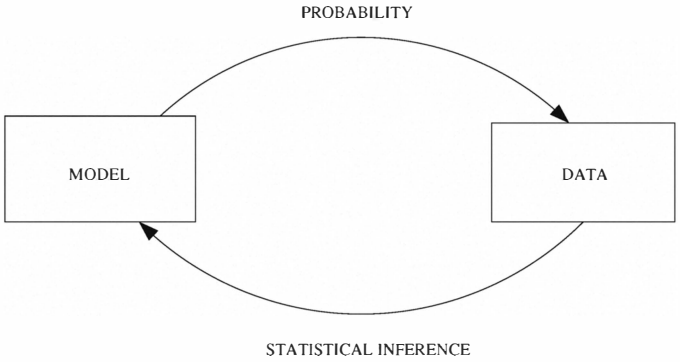

Figure 4.1 An illustration of the dual roles of probability and statistics in data analysis. Probability specifies how observed data can be generated from models. Statistical inference allows us to infer models from observed data.

It is when a sample is used that the power of inferential statistics comes into play. *Statistical inference* allows us to make statements about population structures, to estimate the size of these structures, and to state our degree of confidence in them, all on the basis of a sample. (See figure 4.1 for a simple illustration of the roles of probability and statistics). Thus, for example, we could say that our best estimate of a population value is 6.3, and that one is 95% confident that the true population value lies between 5.9 and 6.7. (Definition and interpretation of intervals such as these is a delicate point, and depends on what philosophical basis we adopt—frequentist or Bayesian, for example. We shall say more about such intervals later in this chapter.) Note the use of the word *estimate* for the population value here. If we were basing our analysis on the entire population, we would use the word *calculate*: if all the constituent numbers are known, we can actually calculate the population value, and no notion of estimation arises.

In order to make an inference about a population structure, we must have a model or pattern structure in mind: we would not be able to assess the evidence for some structure underlying the data if we never contemplated the existence of such a structure. So, for example, we might hypothesize that the value of some variable Z depends on the values of two other variables X and Y. Our model is that Z is related to X and Y. Then we can estimate the strength of these relationships in the population. (Of course, we may conclude that one or both of the relationships are of strength zero—that there

is no relationship.)

Statistical inference is based on the premise that the sample has been drawn from the population in a random manner—that each member of the population had a particular probability of appearing in the sample. The model will specify the distribution function for the population—the probability that a particular value for the random variable will arise in the sample. For example, if the model indicates that the data have arisen from a Normal distribution with a mean of 0 and a standard deviation of 1, it also tells us that the probability of observing a value as large as +20 is very small. Indeed, under the assumption that the model is correct, a precise probability can be put on observing a value greater than +20. Given the model, we can generally compute the probability that an observation will fall within any interval. For samples from categorical distributions, we can estimate the probability that values equal to each of the observed values would have arisen. In general, if we have a model M for the data we can state the probability that a random sampling process would lead to the data $D = \{\mathbf{x}(1), \ldots, \mathbf{x}(n)\}$, here $\mathbf{x}(i)$ is the ith p-dimensional vector of measurements (the ith row in our $n \times p$ data matrix). This probability is expressed as $p(D \mid M)$. Often we do not make dependence on the model M explicit and simply write $p(D)$, relying on the context to make it clear. (As noted in the appendix the probability of observing any *particular* value of a variable that has a continuous cumulative distribution function is zero—particular values refer to intervals of length zero, and therefore the area under the probability density function across such an interval is zero. However, all real data actually refer to finite (if small) intervals (e.g., if someone is said to be 5 feet 11 inches tall, they are known to have a height in the interval between 5 feet 10.5 inches and 5 feet 11.5 inches). Thus it does make sense to talk of the probability of any particular data value being observed in practice.)

Let $p(\mathbf{x}(i))$ be the probability of individual i having vector measurement $\mathbf{x}(i)$ (here p could be a probability mass function or a density function, depending on the nature of $\mathbf{x}$). If we further assume that the probability of each member of the population being selected for inclusion in the sample has no effect on the probability of other members being selected (that is, that the separate observations are independent, or that the data are drawn "at random"), the overall probability of observing the entire distribution of values in the sample is simply the product of the individual probabilities:

$$p(D \mid \theta, M) = \prod_{i=1}^{n} p\left(\mathbf{x}(i) \mid \theta, M\right), \tag{4.3}$$

where M is the model and θ are the parameters of the model (assumed fixed at this point). (When regarded as a function of the parameters θ in the model M, this is called the *likelihood function*. We discuss it in detail below.) Methods have been developed to cope with situations in which observing one value alters the chance of observing another, but independence is by far the most commonly used assumption, even when it is only approximately true.

Based on this probability, we can decide how realistic the assumed model is. If our calculations suggest it is very unlikely that the assumed model would have given rise to the observed data, we might feel justified in rejecting the model; this is the principle underlying hypothesis tests (section 4.6). In hypothesis testing we decide to reject an assumed model (the null hypothesis) if the probability of the observed data arising under that model is less than some pre-specified value (often 0.01 or 0.05—the *significance level* of the test).

A similar principle is used in estimating population values for the parameters of the model. Suppose that our model indicates that the data arise from a Normal distribution with unit variance but unknown mean μ. We could propose various values for the mean, for each one calculating the probability that the observed data would have arisen if the population mean had that value. We could carry out hypothesis tests for each value, rejecting those with a low probability of having given rise to the observed data. Or we can short-cut this process and simply use the estimate of the mean with the highest probability of having generated the observed data. This value is called the maximum likelihood estimate of the mean, and the process we have described is maximum likelihood estimation (see section 4.5). The probability that a particular model would give rise to the observed data, when expressed as a function of the parameters, is called the likelihood function. This function can also be used to define an interval of likely values; we can say, for example, that, assuming our model is correct, 90% of intervals generated from a data sample in this way will contain the true value of the parameter.

4.5 Estimation

In chapter 3 we described several techniques for summarizing a given set of data. When we are concerned with inference, we want to make more general statements, statements about the entire population of values that *might* have been drawn. These are statements about the probability distribution or probability density function (or, equivalently, about the cumulative distribution

function) from which the data are assumed to have arisen.

4.5.1 Desirable Properties of Estimators

In the following subsections we describe the two most important methods of estimating the parameters of a model: maximum likelihood estimation and Bayesian estimation. It is important to be aware of the differing properties of different methods so that we can adopt a method suited to our problem. Here we briefly describe some attractive properties of *estimators*. Let $\hat{\theta}$ be an estimator of a parameter θ. Since $\hat{\theta}$ is a number derived from the data, if we were to draw a different sample of data, we would obtain a different value for $\hat{\theta}$. Thus, $\hat{\theta}$ is a random variable. Therefore, it has a distribution, with different values arising as different samples are drawn. We can obtain descriptive summaries of that distribution. It will, for example, have a mean or expected value, $E[\hat{\theta}]$. Here the expectation function E is taken with respect to the true (unknown) distribution from which the data are assumed to be sampled—that is, over all possible data sets of size n that could occur weighted by their probability of occurrence.

The *bias* of $\hat{\theta}$ (a concept we introduced informally in chapter 2) is defined as

$$Bias(\theta) = E[\hat{\theta}] - \theta, \tag{4.4}$$

the difference between the expected value of the estimator $E[\hat{\theta}]$ and the true value of the parameter θ. Estimators for which $E[\hat{\theta}] = \theta$ have bias 0 are said to be *unbiased*. Such estimators show no *systematic* departure from the true parameter value on average, although for any particular single data set D we might have that $\hat{\theta}$ is far away from θ. Note that since both the sampling distribution and the true value of θ are unknown in practice, we cannot typically calculate the actual bias for a given data set. Nonetheless, the general concept of bias (and variance, below) is of fundamental importance in estimation.

Just as the bias of an estimator can be used as a measure of its quality, so also can its variance:

$$Var(\hat{\theta}) = E[\hat{\theta} - E[\hat{\theta}]]^2. \tag{4.5}$$

The variance measures the random, data-driven component of error in our estimation procedure; it reflects how sensitive our estimator will be to the idiosyncrasies of individual data sets. Note that the variance does not depend on the true value of θ—it simply measures how much our estimates will vary across different observed data sets. Thus, although the true sampling distribution is unknown, we can in principle get a data-driven estimate

of the variance of an estimator, for a given value of n, by repeatedly subsampling our original data set and calculating the variance of the estimated $\hat{\theta}$s across these simulated samples. We can choose between estimators that have the same bias by choosing one with minimum variance. Unbiased estimators that have minimum variance are called, unsurprisingly, *best unbiased estimators*.

As an extreme example, if we were to completely ignore our data D and simply say arbitrarily that $\hat{\theta} = 1$ for every data set, then $var(\hat{\theta})$ is zero since the estimate $\hat{\theta}$ never changes as D changes—however this would be a very ineffective estimator in practice since unless we made a very lucky guess we are almost certainly wrong in our estimate of θ, i.e., there will be a non-zero (and potentially very large) bias.

The *mean squared error* of $\hat{\theta}$ is $E[(\hat{\theta}-\theta)^2]$, the mean of the squared difference between the value of the estimator and the true value of the parameter. Mean squared error has a natural decomposition as the sum of the squared bias of $\hat{\theta}$ and its variance:

$$\begin{aligned} E\left[(\hat{\theta}-\theta)^2\right] &= E\left[(\hat{\theta}-E[\hat{\theta}]+E[\hat{\theta}]-\theta)^2\right] \\ &= \left(E[\hat{\theta}]-\theta\right)^2 + E\left[(\hat{\theta}-E[\hat{\theta}])^2\right] \\ &= \left(Bias(\hat{\theta})\right)^2 + Var(\hat{\theta}), \end{aligned} \tag{4.6}$$

where in going from the first to second lines above we took advantage of the fact that various cross-terms in the squared expression cancel out, noting (for example) that $E[\theta] = \theta$ since θ is a constant, etc. Mean squared error is a very useful criterion since it incorporates both systematic (bias) and random (variance) differences between the estimated and true values. (Of course it too is primarily of theoretical interest, since to calculate it we need to know θ, which we don't in practice). Unfortunately, bias and variance often work in different directions: modifying an estimator to reduce its bias increases its variance, and vice versa. The trick is to arrive at the best compromise. Balancing bias and variance is a central issue in data mining and we will return to this point in chapter 6 in a general context and in later chapters in more specific contexts.

There are also more subtle aspects to the use of mean squared error in estimation. For example, mean squared error treats equally large departures from θ as equally serious, regardless of whether they are above or below θ. This is appropriate for measures of location, but may not be appropriate for

measures of dispersion (which, by definition, have a lower bound of zero) or for estimates of probabilities or probability densities.

Suppose that we have a sequence $\hat{\theta}_{n_1}, \ldots, \hat{\theta}_{n_m}$ of estimators, based on increasing sample sizes $n_1 \ldots, n_m$. The sequence is said to be *consistent* if the probability of the difference between $\hat{\theta}$ and the true value θ being greater than any given value tends to 0 as the sample size increases. This is clearly an attractive property (especially in data mining contexts, with large samples) since the larger the sample is the closer the estimator is likely to be to the true value (assuming that the data are coming from a particular distribution—as discussed in chapters 1 and 2, for very large databases this may not be a reasonable assumption).

4.5.2 Maximum Likelihood Estimation

Maximum likelihood estimation is the most widely used method of parameter estimation. Consider a data set of n observations $D = \{\mathbf{x}, \ldots, \mathbf{x}(n)\}$, independently sampled from the same distribution $f(\mathbf{x} \mid \theta)$ (as statisticians say, *independently and identically distributed* or *iid*). The *likelihood function* $L(\theta \mid \mathbf{x}(1), \ldots, \mathbf{x}(n))$ is the probability that the data would have arisen, for a given value of θ, regarded as a function of θ, i.e., $p(D \mid \theta)$. Note that although we are implicitly assuming a particular model M here, as defined by $f(\mathbf{x} \mid \theta)$, for convenience we do not explicitly condition on M in our likelihood definitions below—later, when we consider multiple models we will need to explicitly keep track of which model we are talking about.

Since we have assumed that the observations are independent we have

$$\begin{aligned} L(\theta \mid D) &= L(\theta \mid \mathbf{x}(1), \ldots, \mathbf{x}(n)) \\ &= p(\mathbf{x}(1), \ldots, \mathbf{x}(n) \mid \theta) \\ &= \prod_{i=1}^{n} f(\mathbf{x}(i) \mid \theta), \end{aligned} \tag{4.7}$$

which is a scalar function of θ (where θ itself may be a vector of parameters rather than a single parameter). The *likelihood of a data set* $L(\theta \mid D)$, the probability of the actual observed data D for a particular model, is a fundamental concept in data analysis. Defining a likelihood for a given problem amounts to specifying a probabilistic model for how the data were generated. It turns out that once we can state such a likelihood, the door is opened to the application of many general and powerful ideas from statistical inference. Note that since likelihood is defined as a function of θ the convention is that we

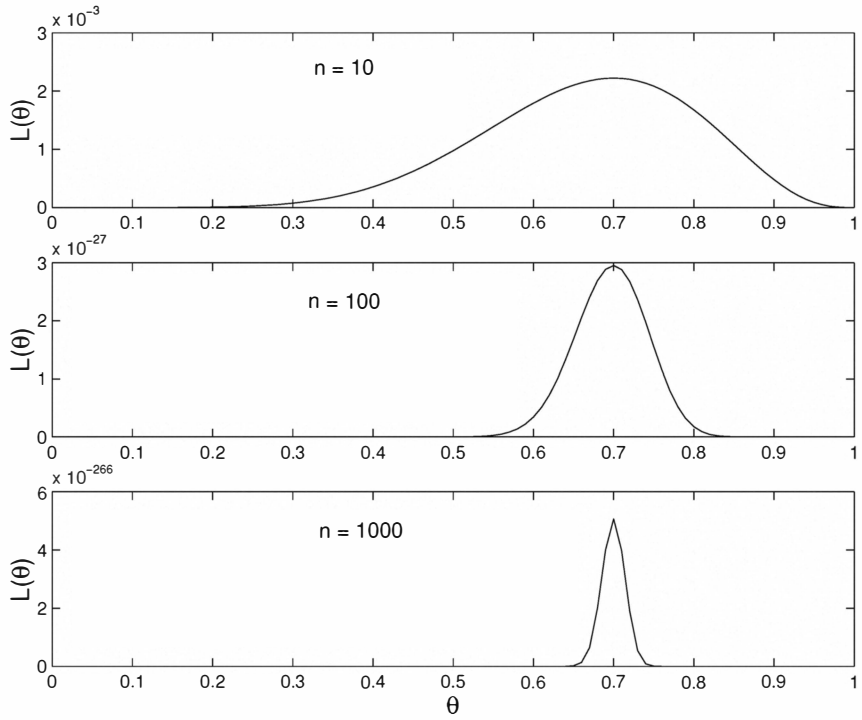

Figure 4.2 The likelihood function for three hypothetical data sets under a Binomial model: $r = 7, n = 10$ (top), $r = 70, n = 100$ (center), and $r = 700, n = 1000$ (bottom).

can drop or ignore any terms in $p(D \mid \theta)$ that do not contain θ, i.e., likelihood is only defined within an arbitrary scaling constant, so it is the shape as a function of θ that matters and not the actual values that it takes. Note also that the idd assumption above is not necessary to define a likelihood: for example, if our n observations had a Markov dependence (where each $\mathbf{x}(i)$ depends on $\mathbf{x}(i-1)$, we would define the likelihood as a product of terms such as $f(\mathbf{x}(i) \mid \mathbf{x}(i-1), \theta)$.

The value for θ for which the data has the highest probability of having arisen is the *maximum likelihood estimator* (or MLE). We will denote the maximum likelihood estimator for θ as $\hat{\theta}_{ML}$.

Example 4.4 Customers in a supermarket either purchase or do not purchase milk. Suppose we want an estimate of the proportion of customers purchasing

milk, based on a sample $x(1), \ldots, x(1000)$ of 1000 randomly drawn observations from the database. Here $x(i)$ takes the value 1 if the ith customer in the sample does purchase milk and 0 if he or she does not. A simple model here would be the observations independently follow a Binomial distribution (described in the appendix) with unknown parameter $0 \leq \theta \leq 1$; that is, θ is the probability that milk is purchased by a random customer. Under the usual assumption of conditional independence given the model, the likelihood can be written as

$$L(\theta \mid x(1), \ldots, x(1000)) = \prod_i \theta^{x(i)}(1-\theta)^{1-x(i)} = \theta^r(1-\theta)^{1000-r},$$

where r is the number among the 1000 who do purchase milk. Taking logs of this yields

$$l(\theta) = \log L(\theta) = r \log \theta + (1000 - r)\log(1-\theta),$$

which, after differentiating and setting to zero, yields

$$\frac{r}{\theta} - \left(\frac{1000-r}{1-\theta}\right) = 0,$$

from which we obtain $\hat{\theta}_{ML} = r/1000$. Thus, the proportion purchasing milk is in fact also the maximum-likelihood estimate of θ under this Binomial model.

In figure 4.2 we plot the likelihood as a function of θ for three hypothetical data sets under this Binomial model. The data sets correspond to 7 milk purchases, 70 milk purchases, and 700 milk purchases out of $n = 10$, $n = 100$, and $n = 1000$, total purchases respectively. The peak of the likelihood function is at the same value, $\theta = 0.7$ in each case, but the uncertainty about the true value of θ (as reflected in the "spread" of the likelihood function) becomes much smaller as n increases (i.e., as we obtain a large customer database). Note that the absolute value of the likelihood function is not relevant; only its shape is of importance.

Example 4.5 Suppose we have assumed that our sample $x(1), \ldots, x(n)$ of n data points has arisen independently from a Normal distribution with unit variance and unknown mean θ. This sort of situation can arise when the source of uncertainty is measurement error; we may know that the results have a certain variance (here rescaled to 1), but not know the mean value for the object that is being repeatedly measured. Then the likelihood function for θ is

$$\begin{aligned} L(\theta \mid x(1), \ldots, x(n)) &= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(-\frac{1}{2}(x(i)-\theta)^2\right) \\ &= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x(i)-\theta)^2\right), \end{aligned}$$

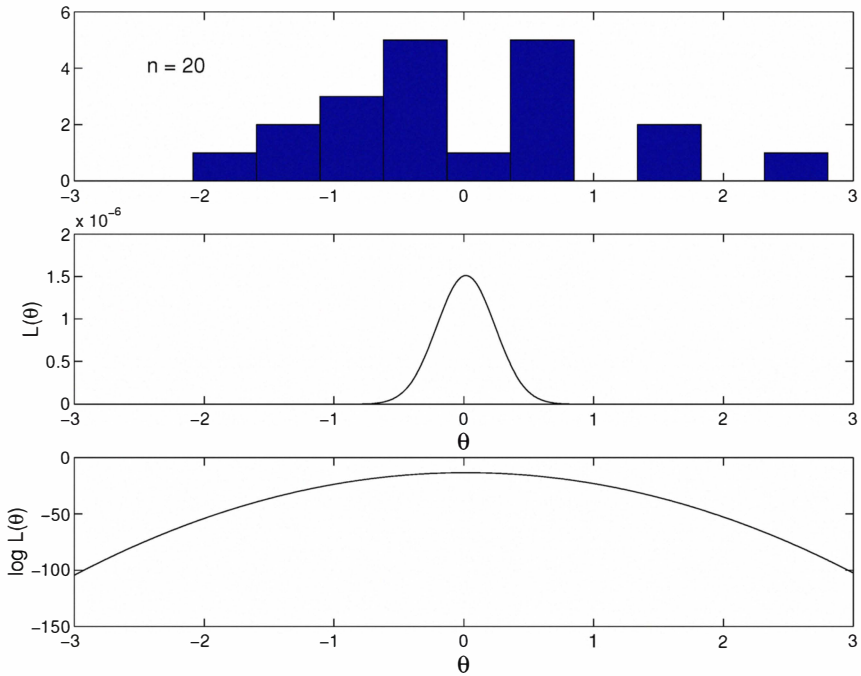

Figure 4.3 The likelihood as a function of θ for a sample of 20 data points from a Normal density with a true mean of 0 and a known standard deviation of 1: (a) a histogram of 20 data points generated from the true model (top), (b) the likelihood function for θ (center), and (c) the log-likelihood function for θ (bottom).

with log-likelihood defined as

$$l(\theta \mid x(1), \ldots, x(n)) = -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{n} (x(i) - \theta)^2, \tag{4.8}$$

To find the MLE we set the derivative $\frac{d}{d\theta} l(\theta \mid x(1), \ldots, x(n))$ to 0 and get

$$\sum_{i=1}^{n} (x(i) - \theta) = 0.$$

Hence, the maximum likelihood estimator $\hat{\theta}_{ML}$ for θ is $\hat{\theta}_{ML} = \sum_i x(i)/n$, the sample mean.

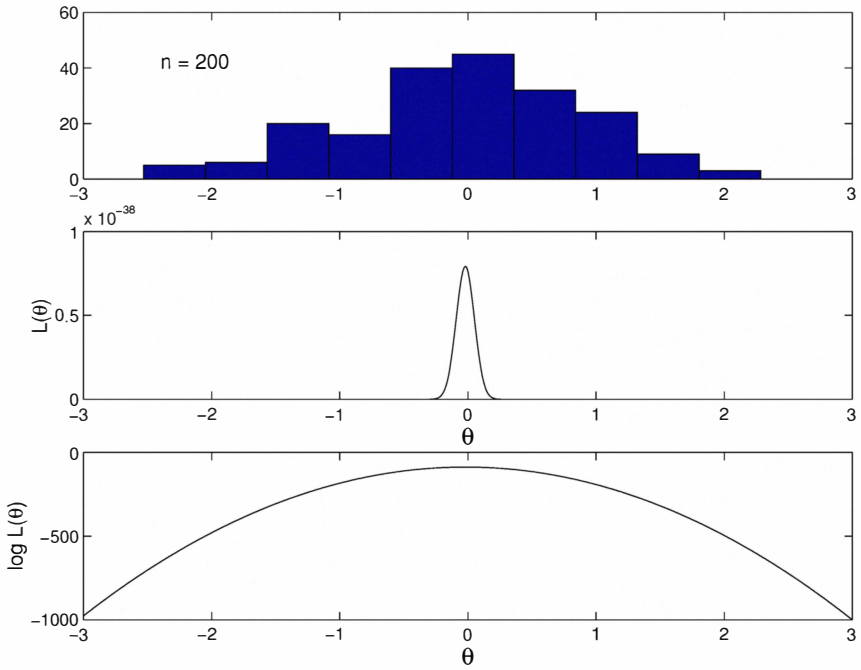

Figure 4.4 The likelihood function for the same model as in figure 4.3 but with 200 data points: (a) a histogram of the 200 data points generated from the true model (top), (b) the likelihood function for θ (center), and (c) the log-likelihood function for θ (bottom).

Figure 4.3 shows both the likelihood function $L(\theta)$ and the log-likelihood $l(\theta) = \log L(\theta)$ as a function of θ for a sample of 20 data points from a Normal density with a true mean of 0 and a known standard deviation of 1. Figure 4.4 shows the same type of plot but with 200 data points. Note how the likelihood function is peaked around the value of the true mean at 0. Also note (as in the Binomial example) how the likelihood function narrows as more data becomes available, reflecting decreasing support from the data for values of θ that are not close to 0.

Example 4.6 A useful general concept in statistical estimation is the notion of a *sufficient statistic*. Loosely speaking, we can define a quantity $s(D)$ as a sufficient statistic for θ if the likelihood $L(\theta)$ only depends on the data through $s(D)$. Thus, in the Binomial model above, the total number of "successes" r

(the number of people who purchase milk) is a sufficient statistic for the Binomial parameter θ. It is sufficient in the sense that the likelihood is only a function of r (assuming n is known already). Knowing which particular customers purchased milk (which particular rows in the data matrix have 1's in the milk column) is irrelevant from the point of view of our Binomial model, once we know the sum total r. Similarly, for the example above involving the estimation of the mean of a Normal distribution, the sum of the observations $\sum_{i=1}^{n} x(i)$ is a sufficient statistic for the likelihood of the mean (keeping in mind that the likelihood is only defined as a function of θ and all other terms can be dropped).

For massive data sets this idea of sufficient statistics can be quite useful in practice—instead of working with the full data set we can simply compute and store the sufficient statistics, knowing that these are sufficient for likelihood-based estimation. For example, if we are gathering large volumes of data on a daily basis (e.g., Web logs) we can in principle just update the sufficient statistics nightly and throw the raw data away. Unfortunately, however, sufficient statistics often do not exist for many of the more flexible model forms that we like to use in data mining applications, such as trees, mixture models, and so forth, that are discussed in detail later in this book. Nonetheless, for simpler models, sufficient statistics are a very useful concept.

Maximum likelihood estimators are intuitively and mathematically attractive; for example, they are consistent estimators in the sense defined earlier. Moreover, if $\hat{\theta}_{ML}$ is the MLE of a parameter θ, then $g(\hat{\theta}_{ML})$ is the MLE of the function $g(\theta)$, though some care needs to be exercised if g is not a one-to-one function. On the other hand, nothing is perfect—maximum likelihood estimators are often biased (depending on the parameter and the underlying model), although this bias may be extremely small for large data sets, often scaling as $O(1/n)$.

For simple problems (where "simple" refers to the mathematical structure of the problem, and not to the number of data points, which can be large), MLEs can be found using differential calculus. In practice, the log-likelihood $l(\theta)$ is usually maximized (as in the Binomial and Normal density examples above), since this replaces the awkward product in the definition with a sum; this process leads to the same result as maximizing $L(\theta)$ directly because the logarithm is a monotonic function. Of course we are often interested in models that have more than one parameter (models such as neural networks (chapter 11) can have hundreds or thousands of parameters). The univariate definition of likelihood generalizes directly to the multivariate case, but in this situation the likelihood is a *mulutivariate function* of d parameters (that

is, a scalar-valued function defined on a d-dimensional parameter space). Since d can be large, finding the maximum of this d-dimensional function can be quite challenging if no closed-form solution exists. We will return to this topic of *optimization* in detail in chapter 8 where we discuss iterative search methods. Multiple maxima can present a difficult problem (which is why stochastic optimization methods are often necessary), as can situations in which optima occur at the boundaries of the parameter space.

Example 4.7 *Simple linear regression* is widely used in data mining. This was mentioned briefly in chapter 1 and is discussed again in detail in chapter 11. In its simplest form it relates two variables: X, a *predictor* or *explanatory* variable, and Y, a *response* variable. The relationship is assumed to take the form $Y = a + bX + e$, where a and b are parameters and e is a random variable assumed to come from a Normal distribution with a mean of 0 and a variance of σ^2, and we can write $e = Y - (a + bX)$. Here the data consists of a set of pairs $D = \{(x(1), y(1)), \ldots, (x(n), y(n))\}$ and the probability density function of the response data given the explanatory data is $f(y(1), \ldots, y(n) \mid x(1), \ldots, x(n), a, b)$. We are interested not in modeling the distribution of the xs, but rather in modeling $f(y|x)$.

Thus, the likelihood (or more precisely, conditional likelihood) function for this model can be written as

$$\begin{aligned} L(a, b|D) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-0.5(y(i) - (a + bx(i))/\sigma)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-0.5/\sigma^2 \sum_{i=1}^{n} (y(i) - (a + bx(i)))^2\right). \end{aligned}$$

To find the maximum likelihood estimators of a and b, we can take logs and discard terms that do not involve either a or b. This yields

$$\sum_{i}^{n} (y(i) - (a + bx(i)))^2 .$$

Thus, we can estimate a and b by finding those values that minimize the sum of squared differences between the predicted values $a + bx(i)$ and the observed values $y(i)$. Such a procedure—minimizing a sum of squares—is ubiquitous in data mining, and goes under the name of the *least squares method*. The sum of squares criterion is of great historical importance, with roots going back to Gauss and beyond. At first it might seem arbitrary to choose a sum of squares (why not a sum of absolute values, for example?), but the above shows how the least squares choice arises naturally from the choice of a Normal distribution for the error term in the model.

Up to this now we have been discussing *point estimates*, single number estimates of the parameter in question. A point estimate is "best" in some sense, but it conveys no idea of the uncertainty associated with it—perhaps there was a large number of almost equally good estimates, or perhaps this estimate was by far the best. Interval estimates provide this sort of information. In place of a single number they give an interval with a specified degree of confidence that this interval contains the unknown parameter. Such an interval is called a *confidence interval*, and the upper and lower limits of the interval are called *confidence limits*. Interpretation of confidence intervals is rather subtle. Here, since we are assuming that θ is unknown but fixed, it does not make sense to say that θ has a certain probability of lying within a given interval: it either does or it does not. However, it does make sense to say that an interval calculated by the given procedure contains θ with a certain probability: after all, the interval is calculated from the sample, and is thus a random variable.

Example 4.8 The following example is deliberately artificial to keep the explanation simple. Suppose the data consist of 100 independent observations from a Normal distribution with unknown mean μ and known variance σ^2, and we want a 95% confidence interval for μ. That is, given the data $x(1), \ldots, x(n)$, we want to find a lower limit $l(x)$ and an upper limit $u(x)$ such that $P(\mu \in [l(x), u(x)]) = 0.95$.

The distribution of the sample mean $\bar{x}$ in this situation (which is also the maximum likelihood estimate of the mean, $\hat{\mu}_{ML}$) is known to follow a Normal distribution with a mean of μ and a variance of $\sigma^2/100$, and hence standard deviation of $\sigma/10$. We also know, from the properties of the Normal distribution (see the appendix), that 95% of the probability lies within 1.96 standard deviations of the mean. Hence,

$$P(\mu - 1.96\sigma/10 \leq \bar{x} \leq \mu + 1.96\sigma/10) = 0.95.$$

This can be rewritten as

$$P(\bar{x} - 1.96\sigma/10 \leq \mu \leq \bar{x} + 1.96\sigma/10) = 0.95.$$

Thus, $l(x) = \bar{x} - 1.96\sigma/10$ and $u(x) = \bar{x} + 1.96\sigma/10$ define a suitable 95% confidence interval.

Frequently confidence intervals are based on the assumption that the sample statistic has a roughly Normal distribution. This is often realistic: the *central limit theorem* tells us that the distribution of many statistics can be approximated well by a Normal distribution, especially if the sample size is

large. Using this approximation, we find an interval in which the statistic has a known probability of lying, given the unknown true parameter value, θ, and invert it to find an interval for the unknown parameter. In order to apply this approach, we need an estimate of the standard deviation of the estimator $\hat{\theta}$. One way to derive such an estimate is the *bootstrap* method.

Example 4.9

Many bootstrap methods, of gradually increasing sophistication and complexity, have been developed over the last two decades. The basic idea is as follows. The data originally arose from a distribution $F(X)$, and we wish to make some statement about this distribution. However, we have only a sample of data $(x(1), ..., x(n))$, which we may denote by $\hat{F}(X)$. What we do is draw a subsample, $\check{F}(X)$, from $\hat{F}(X)$, and act as if $\hat{F}(X)$ were the real distribution. We can repeat this many times, computing a statistic for each of these subsamples. This process gives us information on the sampling properties of statistics calculated from samples drawn from $\hat{F}(X)$, which we hope are similar to the sampling properties of statistics calculated from samples drawn from $F(X)$.

To illustrate, consider an early approach to estimating the performance of a predictive classification rule. As we have discussed above, evaluating performance of a classification rule simply by reclassifying the data used to design it is unwise—it is likely to lead to optimistically biased estimates. Suppose that e_A is the estimate of misclassification rate obtained by the simple resubstitution process of estimating the classification error on the same data as was used to estimate the parameters of the classification model. We really want to estimate e_C, the "true" misclassification rate which we expect to achieve on future objects. The difference between these is $(e_C - e_A)$. If we could estimate this difference, we could adjust e_A to yield a better estimate. In fact, we can estimate this difference, as follows. Suppose we regard $\hat{F}(X)$ as the true distribution and draw from it a subsample—$\check{F}(X)$. Now, acting as if $\hat{F}(X)$ were the true distribution, we can build a rule based on the data in the subsample $\check{F}(X)$ and apply it both to $\hat{F}(X)$ and to $\check{F}(X)$. The difference in performance in these two situations will give us an estimate of the difference $(e_C - e_A)$. To reduce any effects arising from the randomness of the sampling procedure, we repeat the subsampling many times and average the results. The final result is an estimate of the difference $(e_C - e_A)$ that can be added to the value of e_A obtained by resubstituting the data $\hat{F}(X)$ into the rule based on $\hat{F}(X)$, to yield an estimate of the true misclassification rate e_C.

4.5.3 Bayesian Estimation

In the frequentist approach to inference described so far the parameters of a population are fixed but unknown, and the data comprise a random sam-

ple from that population (since the sample was drawn in a random way). The intrinsic variability thus lies in the data $D = \{\mathbf{x}(1), \ldots, \mathbf{x}(n)\}$. In contrast, Bayesian statistics treats the data as known—after all, they have been observed and recorded—and the parameters θ as random variables. Thus, whereas frequentists regard a parameter θ as a fixed but unknown quantity, Bayesians regard θ as having a distribution of possible values and see the observed data as possibly shedding light on this distribution. $p(\theta)$ reflects our degree of belief on where the true (unknown) parameters θ may be. If $p(\theta)$ is very peaked about some value of θ then we are very sure about our convictions (although of course we may be entirely wrong!). If $p(\theta)$ is very broad and flat (and this is the more typical case) then we are expressing a prior belief that is less certain on the location of θ.

Note that while the term *Bayesian* has a fairly precise meaning in statistics, it has sometimes been used in a somewhat looser manner in the computer science and pattern recognition literature to refer to the use of any form of probabilistic model in data analysis. In this text we adopt the more standard and widespread statistical definition, which is described below.

Before the data are analyzed, the distribution of the probabilities that θ will take different values is known as the *prior* distribution $p(\theta)$. Analysis of the data D leads to modification of this distribution to take into account the information in the empirical data, yielding the *posterior* distribution, $p(\theta \mid D)$. The modification from prior to posterior is carried out by means of a theorem named after Thomas Bayes:

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)} = \frac{p(D \mid \theta)p(\theta)}{\int_{\psi} p(D \mid \psi)p(\psi)d\psi}. \tag{4.9}$$

Note that this updating procedure leads to a *distribution*, rather than a single value, for θ. However, the distribution can be used to yield a single value estimate. We could, for example, take the mean of the posterior distribution, or its mode (the latter technique is known as the *maximum a posteriori* method, or MAP). If we choose the prior $p(\theta)$ in a specific manner (e.g., $p(\theta)$ is uniform over some range), the MAP and maximum likelihood estimates of θ may well coincide (since in effect the prior is "flat" and prefers no one value of θ over any other). In this sense, maximum likelihood can be viewed as a special case of the MAP procedure, which in turn is a restricted ("point estimate") form of Bayesian estimation.

For a given set of data D and a particular model, the denominator in equa-

tion 4.9 is a constant, so we can alternatively write the expression as

$$p(\theta \mid D) \propto p(D \mid \theta)p(\theta). \tag{4.10}$$

Here we see that the posterior distribution of θ given D (that is, the distribution conditional on having observed the data D) is proportional to the product of the prior $p(\theta)$ and the likelihood $p(D \mid \theta)$. If we have only weak beliefs about the likely value of the parameter before collecting the data, we will want to choose a prior that spreads the probability widely (for example, a Normal distribution with large variance). In any case, the larger the set of observed data, the more the likelihood dominates the posterior distribution, and the lower the importance of the particular shape of the prior.

Example 4.10 Consider example 4.4 once again involving the proportion of customers who purchase milk, where we consider a single binary variable X and wish to estimate $\theta = p(X = 1)$. A widely used prior for a parameter θ that varies between 0 and 1 is the *Beta* distribution, defined as

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{4.11}$$

where $\alpha > 0, \beta > 0$ are the two parameters of this model. It is straightforward to show that $E[\theta] = \frac{\alpha}{\alpha+\beta}$, that the mode of θ is $\frac{\alpha-1}{\alpha+\beta-2}$, and the variance is $var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Thus, if we assume for example that α and β are chosen to be both greater than 1, we can see that the relative sizes of α and β control the location of both the mean and the mode: if $\alpha = \beta$ then the mean and the mode are at 0. If $\alpha < \beta$ then the mode is less than 0.5, and so forth.

Similarly, the variance is inversely proportional to $\alpha + \beta$: the size of the sum $\alpha+\beta$ controls the "narrowness" of the prior $p(\theta)$. If α and β are relatively large, we will have a relatively narrow peaked prior about the mode. In this manner, we can choose α and β to reflect any prior beliefs we might have about the parameter θ.

Recall from example 4.4 that the likelihood function for θ under the Binomial model can be written as

$$L(\theta \mid D) = \theta^r(1-\theta)^{n-r}, \tag{4.12}$$

where r is the number of 1's in the n total observations. We see that the Beta and Binomial likelihoods are similar in form: the Beta looks like a Binomial likelihood with $\alpha - 1$ prior successes and $\beta - 1$ prior failures. Thus, in effect, we can think of $\alpha + \beta - 2$ as the equivalent sample size for the prior, i.e., it is as if our Beta prior is based on this many prior observations.

Combining the likelihood and the prior, we get

$$\begin{aligned} p(\theta|D) &\propto p(D|\theta)p(\theta) \\ &= \theta^r(1-\theta)^{n-r}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{r+\alpha-1}(1-\theta)^{n-r+\beta-1}. \end{aligned} \tag{4.13}$$

This is conventiently in the form of another Beta distribution, i.e., the posterior on θ, $p(\theta|D)$, is itself another Beta distribution but with parameters $r+\alpha$ and $n-r+\beta$.

Thus, for example, the mean of this posterior distribution $p(\theta|D)$ is $\frac{r+\alpha}{n+\alpha+\beta}$. This is very intuitive. If $\alpha = \beta = 0$ we get the standard MLE of $\frac{r}{n}$. Otherwise, we get a modified estimate, where not all weight is placed on the data alone (on r and n). For example, in data mining practice, it is common to use the heuristic estimate of $\frac{r+1}{n+2}$ for estimates of probabilities, rather than the MLE, corresponding in effect to using a point estimate based on posterior mean and a Beta prior with $\alpha = \beta = 1$. This has the effect of "smoothing" the estimate away from the extreme values of 0 and 1. For example, consider a supermarket where we wanted to estimate the probability of a particular product being purchased, but in the available sample D we had $r = 0$ (perhaps the product is purchased relatively rarely and no-one happened to buy it in the day we drew a sample). The MLE estimate in this case would be 0, whereas the posterior mean would be $\frac{1}{n+2}$, which is close to 0 for large n but allows for a small (but non-zero) probability in the model for that the product is purchased on an average day.

In general, with high-dimensional data sets (i.e., large p) we can anticipate that certain events will not occur in our observed data set D. Rather than committing to the MLE estimate of a probability $\theta = 0$, which is equivalent to stating that the event is impossible according to the model, it is often more prudent to use a Bayesian estimate of the form described here. For the supermarket example, the prior $p(\theta)$ might come from historical data at the same supermarket, or from other stores in the same geographical location. This allows information from other related analyses (in time or space) to be leveraged, and leads to the more general concept of Bayesian hierarchical models (which is somewhat beyond the scope of this text).

One of the primary distinguishing characteristics of the Bayesian approach is the avoidance of so-called *point-estimates* (such as a maximum likelihood estimate of a parameter) in favor of retaining full knowledge of all uncertainty involved in a problem (e.g., calculating a full posterior distribution on θ).

As an example, consider the Bayesian approach to making a prediction about a new data point $\mathbf{x}(n+1)$, a data point not in our training data set D.

Here x might be the value of the Dow-Jones financial index at the daily closing of the stock-market and $n+1$ is one day in the future. Instead of using a point estimate for $\hat{\theta}$ in our model for prediction (as we would in a maximum likelihood or MAP framework), the Bayesian approach is to average over all possible values of θ, weighted by their posterior probability $p(\theta \mid D)$:

$$\begin{aligned} p(\mathbf{x}(n+1) \mid D) &= \int p(\mathbf{x}(n+1), \theta \mid D) d\theta \\ &= \int p(\mathbf{x}(n+1) \mid \theta) p(\theta \mid D) d\theta, \end{aligned} \tag{4.14}$$

since $\mathbf{x}(n+1)$ is conditionally independent of the training data D, given θ, by definition. In fact, we can take this further and also average over different models, using a technique known as Bayesian model averaging. Naturally, all of this averaging can entail considerably more computation than the maximum likelihood approach. This is a primary reason why Bayesian methods have become practical only in recent years (at least for small-scale data sets). For large-scale problems and high-dimensional data, fully Bayesian analysis methods can impose significant computational burdens.

Note that the structure of equations 4.9 and 4.10 enables the distribution to be updated sequentially. For example, after we build a model with data D_1, we can update it with further data D_2:

$$p(\theta \mid D_1, D_2) \propto p(D_2 \mid \theta) p(D_1 \mid \theta) p(\theta). \tag{4.15}$$

This sequential updating property is very attractive for large sets of data, since the result is independent of the order of the data (provided, of course, that D_1 and D_2 are conditionally independent given the underlying model p).

The denominator in equation 4.9, $p(D) = \int_{\psi} p(D \mid \psi) p(\psi) d\psi$, is called the *predictive distribution* of D, and represents our predictions about the value of D. It includes our uncertainty about θ, via the prior $p(\theta)$, and our uncertainty about D when θ is known, via $p(D \mid \theta)$. The predictive distribution changes as new data are observed, and can be useful for model checking: if observed data D have only a small probability according to the predictive distribution, that distribution is unlikely to be correct.

Example 4.11 Suppose we believe that a single data point x comes from a Normal distribution with unknown mean θ and known variance α—that is, $x \sim N(\theta, \alpha)$. Now suppose our prior distribution for θ is $\theta \sim N(\theta_0, \alpha_0)$, with

known θ_0 and α_0. Then

$$\begin{aligned} p(\theta \mid x) &\propto p(x \mid \theta)p(\theta) \\ &= \frac{1}{\sqrt{2\pi\alpha}} \exp(\frac{-1}{2\alpha}(x-\theta)^2) \frac{1}{\sqrt{2\pi\alpha_0}} \exp(\frac{-1}{2\alpha_0}(\theta-\theta_0)^2) \\ &\propto \exp(-\frac{1}{2}\theta^2(1/\alpha_0 + 1/\alpha) + \theta(\theta_0/\alpha_0 + x/\alpha)). \end{aligned}$$

The mathematics here looks horribly complicated (a fairly common occurrence with Bayesian methods), but consider the following reparameterization. Let

$$\alpha_1 = (\alpha_0^{-1} + \alpha^{-1})^{-1}$$

and

$$\theta_1 = \alpha_1(\theta_0/\alpha_0 + x/\alpha).$$

After some algebraic manipulations we get

$$p(\theta \mid x) \propto \exp(-\frac{1}{2}\theta^2/\alpha_1 + \theta\theta_1/\alpha_1) \propto \exp(-\frac{1}{2}(\theta-\theta_1)^2/\alpha_1).$$

Since this is a probability density function for θ, it must integrate to unity. Hence the posterior on θ has the form

$$p(\theta \mid x) = \frac{1}{\sqrt{2\pi\alpha_1}} \exp(-\frac{1}{2}(\theta-\theta_1)^2/\alpha_1).$$

This is a Normal distribution $N(\theta_1, \alpha_1)$. Thus the Normal prior distribution has been updated to yield a Normal posterior distribution and therefore the complicated mathematics can be avoided. Given a Normal prior for the mean and data arising from a Normal distribution as above, we can obtain the posterior merely by computing the updated parameters. Moreover, the updating of the parameters is not as messy as it might at first seem.

Reciprocals of variances are called *precisions*. Here $1/\alpha_1$, the precision of the updated distribution, is simply the sum of the precisions of the prior and the data distributions. This is perfectly reasonable: adding data to the prior should decrease the variance, or increase the precision. Likewise, the updated mean, θ_1, is simply a weighted sum of the prior mean and the datum x, with weights that depend on the precisions of those two values.

When there are n data points, with the situation described above, the posterior is again Normal, now with updated parameter values

$$\alpha_1 = (1/\alpha_0 + n/\alpha)^{-1}$$

and

$$\theta_1 = \alpha_1(\theta_0/\alpha_0 + \bar{x}n/\alpha).$$

The choice of prior distribution can play an important role in Bayesian analysis (more for small samples than for large samples as mentioned earlier). The prior distribution represents our initial belief that the parameter takes different values. The more confident we are that it takes particular values, the more closely the prior will be bunched around those values. The less confident we are, the larger the dispersion of the prior. In the case of a Normal mean, if we had no idea of the true value, we would want to use a prior that gave equal probability to each possible value, i.e., a prior that was perfectly flat or that had infinite variance. This would not correspond to any *proper* density function (which must have some non-zero values and which must integrate to unity). Still, it is sometimes useful to adopt *improper* priors that are uniform throughout the space of the parameter. We can think of such priors as being essentially flat in all regions where the parameter might conceivably occur. Even so, there remains the difficulty that priors that are uniform for a particular parameter are not uniform for a nonlinear transformation of that parameter.

Another issue, which might be seen as either a difficulty or a strength of Bayesian inference, is that priors show an individual's prior belief in the various possible values of a parameter—and individuals differ. It is entirely possible that your prior will differ from mine and therefore we will probably obtain different results from an analysis. In some circumstances this is fine, but in others it is not. One way to overcome this problem is to use a so-called *reference* prior, a prior that is agreed upon by convention. A common form of reference prior is *Jeffrey's* prior. To define this, we first need to define the *Fisher information*:

$$I(\theta \mid \mathbf{x}) = -E\left[\frac{\partial^2 \log L(\theta \mid x)}{\partial \theta^2}\right] \tag{4.16}$$

for a scalar parameter θ—that is, the negative of the expectation of the second derivative of the log-likelihood. Essentially this measures the curvature or flatness of the likelihood function. The flatter a likelihood function is, the less the information it provides about the parameter values. Jeffrey's prior is then defined as

$$p(\theta) \propto \sqrt{I(\theta \mid \mathbf{x})}. \tag{4.17}$$

This is a convenient reference prior since if $\phi = \phi(\theta)$ is some function of θ, this has a prior proportional to $\sqrt{I(\phi \mid \mathbf{x})}$. This means that a consistent prior will result no matter how the parameter is transformed.

The distributions in the examples display began with a Beta or Normal prior and ended with a Beta or Normal posterior. *Conjugate families* of distri-

butions satisfy this property in general: the prior distribution and posterior distribution belong to the same family. The advantage of using conjugate families is that the complicated updating process can be replaced by a simple updating of the parameters.

We have already remarked that it is straightforward to obtain single point estimates from the posterior distribution. Interval estimates are also easy to obtain—integration of the posterior distribution over a region gives the estimated probability that the parameter lies in that region. When a single parameter is involved and the region is an interval, the result is a *credibility interval*. The shortest possible credibility interval is the interval containing a given probability (say 90%) such that the posterior density is highest over the interval. Given that one is prepared to accept the fundamental Bayesian notion that the parameter is a random variable, the interpretation of such intervals is much more straightforward than the interpretation of frequentist confidence intervals.

Of course, it is a rare model that involves only one parameter. Typically models involve several or many parameters. In this case we can find joint posterior distributions for all parameters simultaneously or for individual (sets of) parameters alone. We can also study conditional distributions for some parameters given fixed values of the others. Until recently, Bayesian statistics provided an interesting philosophical viewpoint on inference and induction, but was of little practical value; carrying out the integrations required to obtain marginal distributions of individual parameters from complicated joint distributions was too difficult (only in rare cases could analytic solutions be found, and these often required the imposition of undesirable assumptions). However, in the last 10 years or so this area has experienced something of a revolution. Stochastic estimation methods, based on drawing random samples from the estimated distributions, enable properties of the distributions of the parameters to be estimated and studied. These methods, called Markov chain Monte Carlo (MCMC) methods are discussed again briefly in chapter 8.

It is worth repeating that the primary characteristic of Bayesian statistics lies in its treatment of uncertainty. The Bayesian philosophy is to make all uncertainty explicit in any data analysis, including uncertainty about the estimated parameters as well as any uncertainty about the model. In the maximum likelihood approach, a point estimate of a parameter is often considered the primary goal, but a Bayesian analyst will report a full posterior distribution on the parameter as well as a posterior on model structures. Bayesian prediction consists of taking weighted averages over pa-

rameter values and model structures (where the weights are proportional to the likelihood of the parameter or model given the data, times the prior). In principle, this weighted averaging can provide more accurate predictions than the alternative (and widely used) approach of conditioning on a single model using point estimates of the parameters. However, in practice, the Bayesian approach requires estimation of the averaging weights, which in high-dimensional problems can be difficult. In addition, a weighted average over parameters or models is less likely to be interpretable if description is a primary goal.

4.6 Hypothesis Testing

Although data mining is primarily concerned with looking for unsuspected features in data (as opposed testing specific hypotheses that are formed before we see the data), in practice we often do want to test specific hypotheses (for example, if our data mining algorithm generates a potentially interesting hypothesis that we would like to explore further).

In many situations we want to see whether the data support some idea about the value of a parameter. For example, we might want to know if a new treatment has an effect greater than that of the standard treatment, or if two variables are related in a population. Since we are often unable to measure these for an entire population, we must base our conclusions on a samples. Statistical tools for exploring such hypotheses are called *hypothesis tests*.

4.6.1 Classical Hypothesis Testing

The basic principle of hypothesis tests is as follows. We begin by defining two complementary hypotheses: the *null hypothesis* and the *alternative hypothesis*. Often the null hypothesis is some point value (e.g., that the effect in question has value zero—that there is no treatment difference or regression slope) and the alternative hypothesis is simply the complement of the null hypothesis. Suppose, for example, that we are trying to draw conclusions about a parameter θ. The null hypothesis, denoted by H_0, might state that $\theta = \theta_0$, and the alternative hypothesis (H_1) might state that $\theta \neq \theta_0$. Using the observed data, we calculate a statistic (what form of statistic is best depends on the nature of the hypothesis being tested; examples are given below). The statistic would vary from sample to sample—it would be a random variable. If we assume that the null hypothesis is correct, then we can determine the

expected distribution for the chosen statistic, and the observed value of the statistic would be one point from that distribution. If the observed value were way out in the tail of the distribution, we would have to conclude either that an unlikely event had occurred or that the null hypothesis was not, in fact, true. The more extreme the observed value, the less confidence we would have in the null hypothesis.

We can put numbers on this procedure. Looking at the top tail of the distribution of the statistic (the distribution based on the assumption that the null hypothesis is true), we can find those potential values that, taken together, have a probability of 0.05 of occurring. These are extreme values of the statistic—values that deviate quite substantially from the bulk of the values, assuming the null hypothesis is true. If this extreme observed value did lie in this top region, we could *reject* the null hypothesis "at the 5% level": only 5% of the time would we expect to see a result in this region—as extreme as this—if the null hypothesis were correct. For obvious reasons, this region is called the *rejection region* or *critical region*. Of course, we might not merely be interested in deviations from the null hypothesis in one direction. That is, we might be interested in the lower tail, as well as the upper tail of the distribution. In this case we might define the rejection region as the union of the test statistic values in the lowest 2.5% of the probability distribution and the test statistic values in the uppermost 2.5% of the probability distribution. This would be a *two-tailed* test, as opposed to the previously described *one-tailed* test. The size of the rejection region, known as the *significance level* of the test, can be chosen at will. Common values are 1%, 5%, and 10%.

We can compare different test procedures in terms of their *power*. The power of a test is the probability that it will correctly reject a false null hypothesis. To evaluate the power of a test, we need a specific alternative hypothesis so we can calculate the probability that the test statistic will fall in the rejection region if the alternative hypothesis is true.

A fundamental question is how to find a good test statistic for a particular problem. One strategy is to use the *likelihood ratio*. The likelihood ratio statistic used to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$ is defined as

$$\lambda = \frac{L(\theta_0 \mid D)}{\sup_\psi L(\psi \mid D)}, \tag{4.18}$$

where $D = \{\mathbf{x}(1), \ldots, \mathbf{x}(n)\}$. That is, the ratio of the likelihood when $\theta = \theta_0$ to the largest value of the likelihood when θ is unconstrained. Clearly, the null hypothesis should be rejected when λ is small. This procedure can

easily be generalized to situations in which the null hypothesis is not a point hypothesis but includes a set of possible values for θ.

Example 4.12 Suppose that we have a sample of n points independently drawn from a Normal distribution with unknown mean and unit variance, and that we wish to test the hypothesis that the mean has a value of 0. The likelihood under this (null hypothesis) assumption is

$$L(0 \mid x(1), \ldots, x(n)) = \prod_i p(x(i) \mid 0) = \prod_i \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x(i) - 0)^2\right).$$

The maximum likelihood estimator of the mean of a Normal distribution is the sample mean, so the unconstrained maximum likelihood is

$$L(\mu \mid x(1), \ldots, x(n)) = \prod_i p(x(i) \mid \mu) = \prod_i \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x(i) - \bar{x})^2\right).$$

The ratio of these simplifies to

$$\lambda = \exp\left(-n(\bar{x} - 0)^2/2\right).$$

Therefore, our rejection region is thus $\{\lambda \mid \lambda \leq c\}$ for a suitably chosen value of c. This expression can be rewritten as

$$\bar{x} \geq \sqrt{-\frac{2}{n} \ln c},$$

where $\bar{x} = \frac{1}{n} \sum_i x(i)$ is the sample mean. Thus, the test statistic $\bar{x}$ has to be compared with a constant.

Certain types of tests are used very frequently. These include tests of differences between means, tests to compare variances, and tests to compare an observed distribution with a hypothesized distribution (so-called *goodness-of-fit tests*). The common t-test of the difference between the means of two independent groups is described in the display below. Descriptions of other tests can be found in introductory statistics texts.

Example 4.13 Let $x(1), \ldots, x(n)$ be a sample of n observations randomly drawn from a Normal distribution $N(\mu_x, \sigma^2)$, and let $y(1), \ldots, y(m)$ be an independent sample of m observations randomly drawn from a Normal distribution $N(\mu_y, \sigma^2)$. Suppose we wish to test the hypothesis that the means are equal, $H_0 : \mu_x = \mu_y$. The likelihood ratio statistic under these circumstances reduces to

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2(1/n + 1/m)}},$$

with

$$s = s_x^2 \frac{n-1}{n+m-2} + s_y^2 \frac{m-1}{n+m-2},$$

where

$$s_x^2 = \sum (s - \bar{x})^2 / (n-1)$$

is the estimated variance for the x sample and s_y^2 is the same coefficient for the ys. The quantity s is thus a simple weighted sum of the sample variances of the two samples, and the test statistic is merely the difference between the two sample means adjusted by the estimated standard deviation of that difference. Under the null hypothesis, t follows a t distribution (see the appendix) with $n + m - 2$ degrees of freedom.

Although the two populations being compared here are assumed to be Normal, this test is fairly robust to departures from Normality, especially if the sample sizes and the variances are roughly equal. This test is very widely used.

Example 4.14 Relationships between variables are often of central interest in data mining. At an extreme, we might want to know if two variables are not related at all, so that the distribution of the value taken by one is the same regardless of the value taken by the other. A suitable test for independence of two categorical variables is the chi-squared test. This is essentially a goodness-of-fit test in which the data are compared with a model based on the null hypothesis of independence.

Suppose we have two variables, x and y, with x taking the values x_i, $i = 1, ..., r$ with probabilities $p(x_i)$ and y taking the values y_j, $j = 1, ..., s$ with probabilities $p(y_j)$. Suppose that the joint probabilities are $p(x_i, y_j)$. Then, if x and y are independent, $p(x_i, y_j) = p(x_i)\, p(y_j)$. The data permit us to estimate the distributions $p(x_i)$ and $p(y_j)$ simply by calculating the proportions of the observations that fall at each level of x and the proportions that fall at each level of y. Let the estimate of the probability of the x variable taking value x_i be $n(x_i)/n$ and $n(x_i)/n$ the estimate of the probability of the y variable taking value y_j. Multiplying these together gives us estimates of the probabilities we would expect in each cell, under the independence hypothesis; thus, our estimate of $p(x_i, y_j)$ under the independence assumption is $n(x_i)\, n(y_j)/n^2$. Since there are n observations altogether, this means we would expect, under the null hypothesis, to find $n(x_i)\, n(y_j)/n$ observations in the (x_i, y_j)th cell. For convenience, number the cells sequentially in some order from 1 to t (so $t = r.s$) and let E_k represent the expected number in the kth cell. We can compare this with the observed number in the kth cell, which we shall denote as O_k. Somehow, we need to aggregate this comparison over all t cells. A

suitable aggregation is given by

$$X^2 = \sum_{k=1,t} \frac{(E_k - O_k)^2}{E_k}. \tag{4.19}$$

The squaring here avoids the problem of positive and negative differences canceling out, and the division by E_k prevents large cells dominating the measure. If the null hypothesis of independence is correct, X^2 follows a χ^2 distribution with $(r-1)(s-1)$ degrees of freedom, so that significance levels can either be found from tables or be computed directly.

We illustrate using medical data in which the outcomes of surgical operations (no improvement, partial improvement, and complete improvement) are classified according to the kind of hospital in which they occur ("referral" or "non-referral"). The data are illustrated below, and the question of interest is whether the outcome is independent of hospital type (that is, whether the outcome distribution is the same for both types of hospital).

	Referral	Non-referral
No improvement	43	47
Partial improvement	29	120
Complete improvement	10	118

The total number of patients from referral hospitals is $(43 + 29 + 10) = 82$, and the total number of patients who do not improve at all is $(43 + 47) = 90$. The overall total is 367. From this it follows that the expected number in the top left cell of the table, under the independence assumption, is $82 \times 90/367 = 20.11$. The observed number is 43, so this cell contributes a value of $(20.11 - 43)^2/20.11$ to X^2. Performing similar calculations for each of the six cells, and adding the results yields $X^2 = 49.8$. Comparing this with a χ^2 distribution with $(3-1)(2-1) = 2$ degrees of freedom reveals a very high level of significance, suggesting that the outcome of surgical operations does depend on hospital type.

The hypothesis testing strategy outlined above is based on the assumption that a random sample has been drawn from some distribution, and the aim of the testing is to make a probability statement about a parameter of that distribution. The ultimate objective is to make an inference from the sample to the underlying population of potential values. For obvious reasons, this is sometimes described as the *sampling paradigm*. An alternative strategy is sometimes appropriate, especially when we are not confident that the sample has been obtained though probability sampling (see chapter 2), and therefore

inference to the underlying population is not possible. In such cases, we can still sometimes make a probability statement about some effect under a null hypothesis. Consider, for example, a comparison of a treatment and a control group. We might adopt as our null hypothesis that there is no treatment effect, so the distribution of scores of people who received the treatment should be the same as that of those who did not. If we took a sample of people (possibly not randomly drawn) and randomly assign them to the treatment and control groups, we would expect the difference of mean scores between the groups to be small if the null hypothesis was true. Indeed, under fairly general assumptions, it is not difficult to work out the distribution of the difference between the sample means of the two groups we would expect if there were no treatment effect, and if such difference were just a consequence of an imbalance in the random allocation. We can then explore how unlikely it is that a difference as large or larger than that actually obtained would be seen. Tests based on this principle are termed *randomization tests* or *permutation tests*. Note that they make no statistical inference from the sample to the overall population, but they do enable us to make conditional probability statements about the treatment effects, conditional on the observed values.

Many statistical tests make assumptions about the forms of the population distributions from which the samples are drawn. For example, in the two-sample t-test, illustrated above, an assumption of Normality was made. Often, however, it is inconvenient to make such assumptions. Perhaps we have little justification for the assumption, or perhaps we know that the data do not to follow the form required by a standard test. In such circumstances we can adopt *distribution-free tests*. Tests based on ranks fall into this class. Here the basic data are replaced by the numerical labels of the positions in which they occur. For example, to explore whether two samples arose from the same distribution, we could replace the actual numerical values by their ranks. If they did arise from the same distribution, we would expect the ranks of the members of the two samples to be well mixed. If, however, one distribution had a larger mean than the other, we would expect one sample to tend to have large ranks and the other to have small ranks. If the distributions had the same means but one sample had a larger variance than the other, we would expect one sample to show a surfeit of large and small ranks and the other to dominate the intermediate ranks. Test statistics can be constructed based on the average values or some other measurements of the ranks, and their significance levels can be evaluated using randomization arguments. Such test statistics include the sign test statistic, the rank sum test statistic, the Kolmogorov-Smirnov test statistic, and the Wilcoxon test statis-

tic. Sometimes the term *nonparametric test* is used to describe such tests—the rationale being that these tests are not testing the value of a parameter of any assumed distribution.

Comparison of hypotheses H_0 and H_1 from a Bayesian perspective is achieved by comparing their posterior probabilities:

$$p(H_i|x) \propto p(x|H_i)\, p(H_i) \tag{4.20}$$

Taking the ratio of these leads to a factorization in terms of the prior odds and the likelihood ratio, or *Bayes factor*:

$$\frac{p(H_0|x)}{p(H_1|x)} \propto \frac{p(H_0)}{p(H_1)} . \frac{p(x|H_0)}{p(x|H_1)} . \tag{4.21}$$

There are some complications here, however. The likelihoods are *marginal likelihoods* obtained by integrating over parameters not specified in the hypotheses, and the prior probabilities will be zero if the H_i refer to particular values from a continuum of possible values (e.g., if they refer to values of a parameter θ, where θ can take any value between 0 and 1). One strategy for dealing with this problem is to assign a discrete non-zero prior probability to the given values of θ.

4.6.2 Hypothesis Testing in Context

This section has so far described the classical (frequentist) approach to statistical hypothesis testing. In data mining, however, analyses can become more complicated.

Firstly, because data mining involves large data sets, we should *expect* to obtain statistical significance: even slight departures from the hypothesized model form will be identified as significant, even though they may be of no practical importance. (If they are of practical importance, of course, then well and good.) Worse, slight departures from the model arising from contamination or data distortion will show up as significant. We have already remarked on the inevitability of this problem.

Secondly, sequential model fitting processes are common. Beginning in chapters 8 we will describe various *stepwise* model fitting procedures, which gradually refine a model by adding or deleting terms. Running separate tests on each model, as if it were *de novo*, leads to incorrect probabilities. Formal sequential testing procedures have been developed, but they can be quite complex. Moreover, they may be weak because of the multiple testing going on.

Thirdly, the fact that data mining is essentially an exploratory process has various implications. One is that many models will be examined. Suppose we test m true (though we will not know this) null hypotheses at the 5% level, each based on its own subset of the data, independent of the other tests. For each hypothesis separately, there is a probability of 0.05 of incorrectly rejecting the hypothesis. Since the tests are independent, the probability of incorrectly rejecting at least one is $p = 1 - (1 - 0.05)^m$. When $m = 1$ we have $p = 0.05$, which is fine. But when $m = 10$ we obtain $p = 0.4013$, and when $m = 100$ we obtain $p = 0.9941$. Thus, if we test as few as even 100 true null hypotheses, we are almost certain to incorrectly reject at least one. Alternatively, we could control the overall *family* error rate, setting the probability of incorrectly rejecting one of more of the m true null hypotheses to 0.05. In this case we use $0.05 = 1 - (1 - \alpha)^m$ for each given m to obtain the level α at which each of the separate null hypotheses is tested. With $m = 10$ we obtain $\alpha = 0.0051$, and with $m = 100$ we obtain $\alpha = 0.0005$. This means that we have a very small probability of incorrectly rejecting any of the separate component hypotheses.

Of course, in practice things are much more complicated: the hypotheses are unlikely to be completely independent (at the other extreme, if they are completely dependent, accepting or rejecting one implies the acceptance or rejection of all), with an essentially unknowable dependence structure, and there will typically be a mixture of true (or approximately true) and false null hypotheses.

Various *simultaneous test procedures* have been developed to ease these difficulties (even though the problem is not really one of inadequate methods, but is really more fundamental). A basic approach is based on the *Bonferroni* inequality. We can expand the probability $(1 - \alpha)^m$ that none of the true null hypotheses are rejected to yield $(1 - \alpha)^m \geq 1 - m\alpha$. It follows that $1 - (1 - \alpha)^m \leq m\alpha$—that is, the probability that one or more true null hypotheses is incorrectly rejected is less than or equal to $m\alpha$. In general, the probability of incorrectly rejecting one or more of the true null hypotheses is smaller than the sum of probabilities of incorrectly rejecting each of them. This is a first-order *Bonferroni inequality*. By including other terms in the expansion, we can develop more accurate bounds—though they require knowledge of the dependence relationships between the hypotheses.

With some test procedures difficulties can arise in which a global test of a family of hypotheses rejects the null hypothesis (so we believe at least one to be false), but no single component is rejected. Once again strategies have been developed for overcoming this in particular applications. For example,

in multivariate analysis of variance, which compares several groups of objects that have been measured on multiple variables, test procedures have been developed that overcome these problems by comparing each test statistic with a single threshold value.

It is obvious from the above discussion that while attempts to put probabilities on statements of various kinds, via hypothesis tests, do have a place in data mining, they are not a universal solution. However, they can be regarded as a particular type of a more general procedure that maps the data and statement to a numerical value or *score*. Higher scores (or lower scores, depending upon the procedure) indicate that one statement or model is to be preferred to another, without attempting any absolute probabilistic interpretation. The penalized goodness-of-fit score functions described in chapter 7 can be thought of in this context.

4.7 Sampling Methods

As mentioned earlier, data mining can be characterized as secondary analysis, and data miners are not typically involved directly with the data collection process. Still, if we have information about that process that might be useful for our analysis, we should take advantage of it. Traditional statistical data collection is usually carried out with a view to answering some particular question or questions in an efficient and effective manner. However, since data mining is a process seeking the unexpected or the unforeseen, it does not try to answer questions that were specified before the data were collected. For this reason we will not be discussing the sub-discipline of statistics known as *experimental design*, which is concerned with optimal ways to collect data. The fact that data miners typically have no control over the data collection process may sometimes explain poor data quality: the data may be ideally suited to the purposes for which it was collected, but not adequate for its data mining uses.

We have already noted that when the database comprises the entire population, notions of statistical inference are irrelevant: if we want to know the value of some population parameter (the mean transaction value, say, or the largest transaction value), we can simply calculate it. Of course, this assumes that the data describe the population perfectly, with no measurement error, missing data, data corruption, and so on. Since, as we have seen, this is an unlikely situation, we may still be interested in making an inference from the data as recorded to the "true" underlying population values.

Furthermore, the notions of populations and samples can be deceptive. For example, even when values for the entire population have been captured in the database, often the aim is not to describe that population, but rather to make some statement about likely future values. For example, we may have available the entire population of transactions made in a chain of supermarkets on a given day. We may well wish to make some kind of inferential statement—statement about the mean transaction value for the next day or some other future day. This also involves uncertainty, but it is of a different kind from that discussed above. Essentially, here, we are concerned with *forecasting*. In market basket analysis we do not really wish to describe the purchasing patterns of last month's shoppers, but rather to forecast how next month's shoppers are likely to behave.

We have distinguished two ways in which samples arise in data mining. First, sometimes the database itself is merely a sample from some larger population. In chapter 2 we discussed the implications of this situation and the dangers associated with it. Second the database contains records for every object in the population, but the analysis of the data is based on only a sample from it. This second technique is appropriate only in modeling situations and certain pattern detection situations. It is not appropriate when we are seeking individual unusual records.

Our aim is to draw a sample from the database that allows us to construct a model that reflects the structure of the data in the database. The reason for using just a sample, rather than the entire data set, is one of efficiency. At an extreme, it may be infeasible, in terms of time or computational requirements, to use the entirety of a large database. By basing our computations solely on a sample, we make the computations quicker and easier. It is important, however, that the sample be drawn in such a way that it reflects the structure of the complete set—i.e., that it is representative of the entire database.

There are various strategies for drawing samples to try to ensure representativeness. If we wanted to take just 1 in 2 of the records (a *sampling fraction* of 0.5), we could simply take every other record. Such a direct approach is termed *systematic sampling*. Often it is perfectly adequate. However, it can also lead to unsuspected problems. For instance, if the database contained records of married couples, with husbands and wives alternating, systematic sampling could be disastrous—the conclusions drawn would probably be entirely mistaken. In general, in any sampling scheme in which cases are selected following some regular pattern there is a risk of interaction with an unsuspected regularity in the database. Clearly what we need is a selection

pattern that avoids regularities—a random selection pattern.

The word random is used here in the sense of avoiding regularities. This is slightly different from the usage employed previously in this chapter, where the term referred to the mechanism by which the sample was chosen. There it described the probability that a record would be chosen for the sample. As we have seen, samples that are random in this second sense can be used as the basis for statistical inference: we can, for example, make a statement about how likely it is that the sample mean will differ substantially from the population mean.

If we draw a sample using a random process, the sample will satisfy the second meaning and is likely to satisfy the first as well. (Indeed, if we specify clearly what we mean by "regularities" we can give a precise probability that a randomly selected sample will not match such regularities.) To avoid biasing our conclusions, we should design our sample selection mechanism in such a way that that each record in the database has an equal chance of being chosen. A sample with equal probability of selecting each member of the population is known as an *epsem* sample. The most basic form of epsem sampling is simple random sampling, in which the n records comprising the sample are selected from the N records in the database in such a way that each set of n records has an equal probability of being chosen. The estimate of the population mean from a simple random sample is just the sample mean.

At this point we should note the distinction between sampling with replacement and sampling without replacement. In the former, a record selected for inclusion in the sample has a chance of being drawn again, but in the latter, once a record is drawn it cannot be drawn a second time. In data mining since the sample size is often small relative to the population size, the differences between the results of these two procedures are usually negligible.

Figure 4.5 illustrates the results of a simple random sampling process used in calculating the mean value of a variable for some population. It is based on drawing samples from a population with a true mean of 0.5. A sample of a specified size is randomly drawn and its mean value is calculated; we have repeated this procedure 200 times and plotted histograms of the results. Figure 4.5shows the distribution of sample mean values (a) for samples of size 10, (b) size 100, and (c) size 1000. It is apparent from this figure that the larger the sample, the more closely the values of the sample mean are distributed around about the true mean. In general, if the variance of a population of size N is σ^2, the variance of the mean of a simple random sample of size n

from that population, drawn without replacement, is

$$\frac{\sigma^2}{n}(1 - \frac{n}{N}). \tag{4.22}$$

Since we normally deal with situations in which N is large relative to n (i.e., situations that involve a small sampling fraction), we can usually ignore the second factor, so that, a good approximation of the variance is σ^2/n. From this it follows that the larger the sample is the less likely it is that the sample mean will deviate significantly from the population mean—which explains why the dispersion of the histograms in figure 4.5 decreases with increasing sample size. Note also that this result is independent of the population size. What matters here is the size of the sample, not the size of the sampling fraction, and not the proportion of the population that is included in the sample. We can also see that, when the sample size is doubled, the standard deviation is reduced not by a factor of 2, but only by a factor of $\sqrt{2}$—there are diminishing returns to increasing the sample size. We can estimate σ^2 from the sample using the standard estimator

$$\sum(x(i) - \bar{x})/(n - 1), \tag{4.23}$$

where $x(i)$ is the value of the ith sample unit and $\bar{y}$ is the mean of the n values in the sample.

The simple random sample is the most basic form of sample design, but others have been developed that have desirable properties under different circumstances. Details can be found in books on survey sampling, such as those cited at the end of this chapter. Here we will briefly describe two important schemes.

In *stratified random sampling*, the entire population is split into nonoverlapping subpopulations or strata, and a sample (often, but not necessarily, a simple random sample) is drawn separately from within each stratum. There are several potential advantages to using such a procedure. An obvious one is that it enables us to make statements about each of the subpopulations separately, without relying on chance to ensure that a reasonable number of observations come from each subpopulation. A more subtle, but often more important, advantage is that if the strata are relatively homogeneous in terms of the variable of interest (so that much of the variability between values of the variable is accounted for by differences between strata), the variance of the overall estimate may be smaller than that arising from a simple random sample. To illustrate, one of the credit card companies we work with categorizes transactions into 26 categories: supermarket, travel agent, gas station,

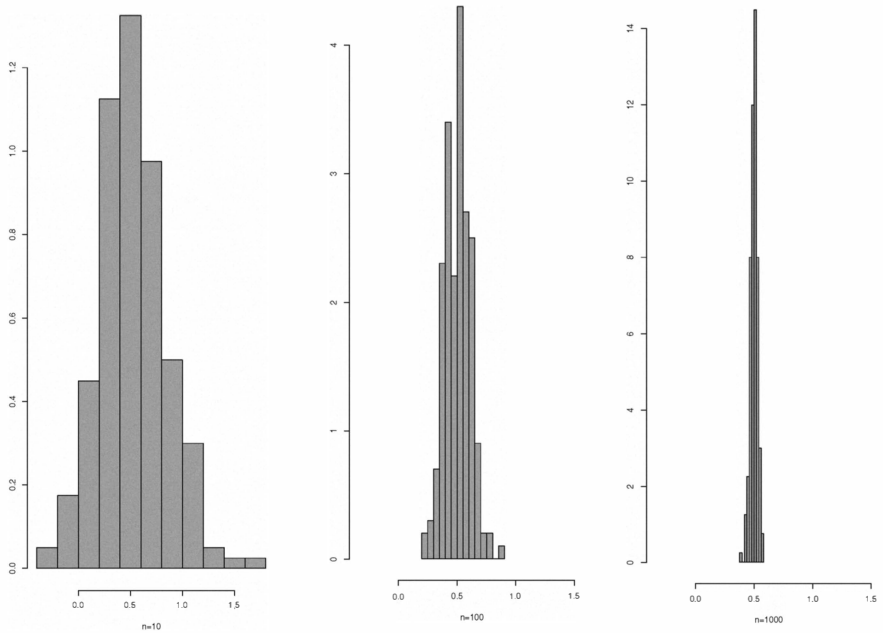

Figure 4.5 Means of samples of size 10(a), 100(b), and 1000(c) drawn from a population with a mean of 0.5.

and so on. Suppose we wanted to estimate the average value of a transaction. We could take a simple random sample of transaction values from the database of records, and compute its mean, using this as our estimate. However, with such a procedure some of the transaction types might end up being underrepresented in our sample, and some might be overrepresented. We could control for this by forcing our sample to include a certain number of each transaction type. This would be a stratified sample, in which the transaction types were the strata. This example illustrates why the strata must be relatively homogeneous internally, with the heterogeneity occurring between strata. If all the strata had the same dispersion as the overall population, no advantage would be gained by stratification.

In general, suppose that we want to estimate the population mean for some variable, and that we are using a stratified sample, with simple ran-

dom sampling within each stratum. Suppose that the kth stratum has N_k elements in it, and that n_k of these are chosen for the sample from this stratum. Denoting the sample mean within the kth stratum by $\bar{x}_k$, the estimate of the overall population mean is given by

$$\sum \frac{N_k \bar{x}_k}{N}, \tag{4.24}$$

where N is the total size of the population. The variance of this estimator is

$$\frac{1}{N^2} \sum N_k^2 var(\bar{x}_k), \tag{4.25}$$

where $var(\bar{x}_k)$ is the variance of the simple random sample of size n_k for the kth stratum, computed as above.

Data often have a hierarchical structure. For example, letters occur in words, which lie in sentences, which are grouped into paragraphs, which occur in chapters, which form books, which sit in libraries. Producing a complete sampling frame and drawing a simple random sample may be difficult. Files will reside on different computers at a site within an organization, and the organization may have many sites; if we are studying the properties of those files, we may find it impossible to produce a complete list from which we can draw a simple random sample. In cluster sampling, rather than drawing a sample of the individual elements that are of interest, we draw a sample of units that contain several elements. In the computer file example, we might draw a sample of computers. We can the examine all of the files on each of the chosen computers, or move on to a further stage of sampling.

Clusters are often of unequal sizes. In the above example we can view a computer as providing a cluster of files, and it is very unlikely that all computers in an organization would have the same number of files. But situations with equal-sized clusters do arise. Manufacturing industries provide many examples: six-packs of beer or packets of condoms, for instance. If all of the units in each selected cluster are chosen (if the subsampling fraction is 1) each unit has the probability a/K of being selected, where a is the number of clusters chosen from the entire set of K clusters. If not all the units are chosen, but the sampling fraction in each cluster is the same, each unit will have the same probability of being included in the sample (it will be an epsem sample). This is a common design. Estimating the variance of a statistic based on such a design is less straightforward than the cases described above since the sample size is also a random variable (it is dependent upon

which clusters happen to be included in the sample). The estimate of the mean of a variable is a ratio of two random variables: the total sum for the units included in the sample and the total number of units included in the sample. Denoting the size of the simple random sample chosen from the kth cluster by n_k, and the total sum for the units chosen from the kth cluster by s_k, the sample mean r is

$$\sum x_k / \sum n_k . \tag{4.26}$$

If we denote the overall sampling fraction by f (often this is small and can be ignored) the variance of r is

$$\frac{1-f}{(\sum n_k)^2} \frac{a}{1-a} \left(\sum s_k^2 + r^2 \sum n_k^2 - 2r \sum s_k n_k \right) . \tag{4.27}$$

4.8 Conclusion

Nothing is certain. In the data mining context, our objective is to make discoveries from data. We want to be as confident as we can that our conclusions are correct, but we often must be satisfied with a conclusion that could be wrong—though it will be better if we can also state our level of confidence in our conclusions. When we are analyzing entire populations, the uncertainty will creep in via less than perfect data quality: some values may be incorrectly recorded, some values may be missing, some members of the population be omitted from the database entirely, and so on. When we are working with samples, our aim is often to draw a conclusion that applies to the broader population from which the sample was drawn. The fundamental tool in tackling all of these issues is probability. This is a universal language for handling uncertainty, a language that has been refined throughout this century and has been applied across a vast array of situations. Application of the ideas of probability enables us to obtain "best" estimates of values, even in the face of data inadequacies, and even when only a sample has been measured. Moreover, application of these ideas also allows us to quantify our confidence in the results.

Later chapters of this book make heavy use of probabilistic arguments. They underlie many—perhaps even most—data mining tools, from global modeling to pattern identification.

4.9 Further Reading

Books containing discussions of different schools of probability, along with the consequences for inference, include those by DeFinetti (1974, 1975), Barnett (1982), and Bernardo and Smith (1994). References to other work on statistics and particular statistical models are given at the ends of chapters 6, 9, 10, and 11.

There are many excellent basic books on the calculus of probability, including those by Grimmett and Stirzaker (1992) and Feller (1968, 1971). The text by Hamming (1991) is oriented towards engineers and computer scientists (and contains many interesting examples), and Applebaum (1996) is geared toward undergraduate mathematics students. Probability calculus is a dynamic area of applied mathematics, and has benefited substantially from the different areas in which it has been applied. For example, Alon and Spencer (1992) give a fascinating tour of the applications of probability in modern computer science.

The idea of randomness as departure from the regular or predictable is discussed in work on Kolmogorov complexity (e.g., Li and Vitanyi, 1993).

Whittaker (1990) provides an excellent treatment of the general principles of conditional dependence and independence in graphical models. Pearl (1988) is a seminal work in this area from the the artificial intelligence perspective.

There are numerous introductory texts on inference, such as those by Daly et al. (1995), as well as more advanced texts that contain a deeper discussion of inferential conscepts, such as Cox and Hinkley (1974), Schervish (1995), Lindsey (1996), and Lehmann and Casella (1998), and Knight (2000). A broad discussion of likelihood and its applications is provided by Edwards (1972). Bayesian methods are now the subjects of entire books. Gelman et al. (1995) provides an excellent general text on Bayesian approach. A comprehensive reference is given by Bernardo and Smith (1994) and a lighter introduction is give by Lee (1989). Nonparametric methods are described by Randles and Wolfe (1979) and Maritz (1981). Bootstrap methods are described by Efron and Tibshirani (1993).

Miller (1980) describes simultaneous test procedures. The methods we have outlined above are not the only approaches to the problem of inference about multiple parameters; Lindsey (1999) describes another.

Books on survey sampling discuss efficient strategies for drawing samples—see, for example, Cochran (1977) and Kish (1965).