CHAPTER 1

Background and Introduction

1.1 Introduction

Statistical analysis is the process of separating out systematic effects from the random noise inherent in all sets of observations. There are three general steps in this process: collection, analysis, and assessment. For most people, data collection is not difficult in that we live in an age where data are omnipresent. More commonly, researchers possess an abundance of data and seek meaningful patterns lurking among the various deadends and distractions. Armed with a substantive theory, many are asking: what should I do now? Furthermore, these same people are often frustrated when receiving multiple, possibly conflicting, answers to that question.

Suppose that there exists a statistical data analysis process with the following desirable characteristics:

- Overt and clear model assumptions.
- A rigorous way to make probability statements about the real quantities of theoretical interest.
- An ability to update these statements (i.e., learn) as new information is received.
- Systematic incorporation of previous knowledge on the subject.
- Missing values handled seamlessly as part of the estimation process.
- Recognition that population quantities are changing over time rather than forever fixed.
- The ability to model a wide class of data types.
- Straightforward assessment of both model quality and sensitivity to assumptions.

As the title of this book suggests, the argument presented here is that

the practice of Bayesian statistics possesses all of these qualities. I will actually argue much beyond this point and assert that the type of data social and behavioral scientists routinely encounter makes the Bayesian approach ideal in ways that traditional data analysis cannot match. These natural advantages include avoiding the assumption of infinite amounts of forthcoming data, recognition that fixed-point assumptions about human behavior are dubious, and a direct way to include existing expertise in the field.

So why do so-called classical approaches dominate Bayesian usage in the social and behavioral sciences? There are several reasons for this. First, key figures in the development of modern statistics had strong prejudices against aspects of Bayesian inference for narrow and subjective reasons. Second, cost of admission is higher in the form of additional mathematical formalism. Third, until recently realistic model specifications sometimes led to unobtainable Bayesian solutions. Finally, there has been a lack of methodological introspection in a number of disciplines. The primary mission of this book is to make the second and third reasons less of a barrier through accessible explanation, detailed examples, and specific guidance on calculation and computing.

1.2 Motivation and Justification

With Bayesian analysis, assertions about unknown model parameters are not expressed in the conventional way as point estimates with reliability assessed using the null hypothesis significance test. Bayesians make no fundamental distinction between observations and unknown parameters are treated as random variables themselves as a logical consequence of Bayesian conditional analysis. Bayesian statistical information about parameters is summarized in probability statements applied to samples or populations in the form of a *posterior distribution*: the distribution of the unknown parameters after observing the data and updating the model. These summary quantities include quantiles of this posterior distribution, the probability of occupying some region of the sample space, the predictive quantities from the posterior, and Bayesian forms of confidence intervals, the credible set and the highest posterior density region.

The essentials of Bayesian thinking are contained in three general steps:

1. Specify a probability model that includes some prior knowledge about the parameters if available for unknown parameter values.

2. Update knowledge about the unknown parameters by conditioning this probability model on observed data.

3. Evaluate the fit of the model to the data and the sensitivity of the conclusions to the assumptions.

Notice that this process does not include the unrealistic and artificial step of making a contrived decision based on some arbitrary quality threshold. The value of a given Bayesian model is instead found in the description of the distribution of some parameter of interest in probabilistic terms. Also, there is nothing about the process contained in the three steps above that cannot be repeated as new data are observed.

There are key assumptions required in the basic Bayesian setup. The first is that a specific parametric form is specified for the unknown parameters.* Secondly, since unknown parameters are treated as having distributional qualities rather than being fixed, it is assumed to be appropriate to specify an initial unconditional distribution on these parameters based on previous substantive knowledge.

Typically (but not always) data values are assumed to be *exchangeable*; the model results are not changed by reordering the data values. This property is more general than, and implied by, the standard assumption that the data are *independent and identically distributed* (iid): independent draws from the same distribution, and also implies a common mean and variance for the data values (Leonard and Hsu 1999, p. 41). Exchangeability allows us to say that the data generation process is conditional on the unknown model parameters in the same way for every data value (de Finetti 1974, Draper et al. 1993, Lindley and Novick 1981). Details about exchangeability are given in Chapter 10.

* *Nonparametric* Bayesian modeling is a large and growing field, but exists beyond the scope of the basic setup.

1.3 Why Are We Uncertain about Probability?

It should be easy to define probability. In fact, it is relatively easy to *mathematically* define the properties of a probability function: (1) it is bounded by zero and one, (2) it sums or integrates to one, and (3) the sum or integral of the probability of disjoint events is equal to the probability of the union of these events (the Kolmogorov axioms (1933), simplified). The real problem lies in describing the actual meaning of probability statements. This difficulty is, in fact, at the heart of traditional disagreements between Bayesians and non-Bayesians.

The frequentist statistical interpretation of probability is that it is a limiting relative frequency: the long-run behavior of a nondeterministic outcome or just an observed proportion in a population. This idea can be traced back to Laplace (1814), who defined probability as the number of successful events out of trials observed. Thus if we could simply repeat the experiment or observe the phenomenon enough times, it would become apparent what the future probability of reoccurrence will be. This is an enormously useful way to think about probability but the drawback is that frequently it is not possible to obtain a large number of outcomes from exactly the same event-generating system (Kendall 1949, Placket 1966).

A competing view of probability is called "subjective" and is often associated with the phrase "degree of belief." Early proponents included Keynes (1921) and Jeffreys (1961), who observed that two people could look at the same situation and assign different probability statements about future occurrences. This perspective is that probability is *personally* defined by the conditions under which a person would make a bet or assume a risk in pursuit of some reward. Subjective probability is closely linked with the idea of decision-making as a field of study (c.f. Bernardo and Smith 1994, Chapter 2) and the principle of selecting choices that maximize personal utility (Berger 1985).

These two characterizations are necessarily simplifications of the perspectives and de Finetti (1974, 1975) provides a much deeper and more detailed categorization. To de Finetti, the ultimate arbiter of subjective probability assignment is the conditions under which individuals will wager their own money. In other words, a person will not violate a personal probability assessment if it has financial consequences. Good (1950) makes this idea

more axiomatic by observing that people have personal probability assessments about many things around them rather than just one, and in order for these disparate comparative statements to form a *body of beliefs* they need to be free of contradictions. For example, if a person thinks that A is more likely to occur than B, and B is more likely to occur than C, then this person cannot coherently believe that C is more likely than A (transitivity). Furthermore, Good adds the explicitly Bayesian idea that people are constantly updating these personal probabilities as new information is observed.

The position underlying nearly all Bayesian work is the subjective probability characterization, although there have been many attempts to "objectify" Bayesian analysis (see Chapter 5). Prior information is formalized in the Bayesian framework and this prior information can be subjective in the sense that the researcher's experience, intuition, and theoretical ideas are included. It is also common to base the prior information on previous studies, experiments, or just personal observations and this process is necessarily subject to a limited (although possibly large) number of observations rather than the infinite number assumed under the frequentist view. We will return to the theme of subjectivity contained in prior information in Chapter 5 and elsewhere, but the principle point is that *all* statistical models include subjective decisions, and therefore we should *ceteris paribus* prefer one that is the most explicit about assumptions. This is exactly the sense that the Bayesian prior provides readers with a specific, formalized statement of currently assumed knowledge in probabilistic terms.

There are some simple but important probability principles and notational conventions that must be understood before proceeding. We will not worry much about the underlying probability theory or measure theory and the concerned reader is directed to the first chapter of any mathematical statistics text or the standard reference works of Billingsley (1986), Chung (1974), and Feller (1990, Volumes 1 and 2). Abstract events are indicated by capital Latin letters: A, B, C, etc. A probability function corresponding to some event A is indicated by $p(A)$. The complement of the event A is denoted A^c, and it is a consequence of Kolmogorov's axioms listed above that $P(A^c) = 1 - P(A)$. The union of two events is indicated by $A \cup B$ and the intersection by $A \cap B$. For any two events: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Two events are independent if $P(A \cap B) = P(A)P(B)$ holds; that is, if the joint distribution can be expressed as the product of the individual marginal distributions.

Central to Bayesian thinking is the idea of conditionality. If an event B is material to another event A in the sense that the occurrence or non-occurrence of B affects the probability of A occurring, then we say that A is conditional on B. It is a basic tenet of Bayesian statistics that we update our probabilities as new relevant information is observed. This is done with the definition of conditional probability given by: $P(A|B) = P(A \cap B)/P(B)$, which is read as "the probability of A given B is equal to the probability of A and B divided by the probability of B."

1.4 Bayes' Law

The Bayesian statistical approach is based on updating information using what is called Bayes' law from his famous 1763 essay.* Like Bayes, Laplace assumed a uniform (equal probability) distribution for the unknown parameter, but he worried much less than Bayes about the consequences of this assumption.

Suppose there are two events of interest A and B, which are not independent. We know from basic axioms of probability that the conditional probability of A given that B has occurred is given by:

$$p(A|B) = \frac{p(A,B)}{p(B)}, \tag{1.1}$$

where $p(A|B)$ is read as "the probability of A given that B has occurred, $p(A,B)$ is the "the probability that both A and B occur," and $p(B)$ is just the unconditional probability that B occurs. Expression (1.1) gives the probability of A after some event B occurs. If A and B are independent then $p(A,B) = p(A)p(B)$ and (1.1) becomes very uninteresting.

* The Reverend Thomas Bayes was an amateur mathematician whose only contribution was an essay found and published two years after his death by his friend Richard Price. The enduring association of an important branch of statistics with his name actually is somewhat of an exaggeration of the generalizability of this work (Stigler 1982). Bayes was the first to explicitly develop this famous law, but it was Laplace (1774, 1781) who (apparently independently) provided a more detailed analysis that is perhaps more relevant to the practice of Bayesian statistics today. See Stigler (1986) for an interesting historical discussion and Sheynin (1977) for a detailed technical analysis.

We can also define a different conditional probability in which A occurs first:

$$p(B|A) = \frac{p(B,A)}{p(A)}. \tag{1.2}$$

Since the probability that A and B occur is the same as the probability that B and A occur ($p(A,B) = p(B,A)$), then we can rearrange (1.1) and (1.2) together in the following way:

$$\begin{aligned} p(A,B) &= p(A|B)p(B) \\ p(B,A) &= p(B|A)p(A) \\ p(A|B)p(B) &= p(B|A)p(A) \\ p(A|B) &= \frac{p(A)}{p(B)}p(B|A). \end{aligned} \tag{1.3}$$

The last line is the famous Bayes' law. This is really a device for "inverting" conditional probabilities. Notice that we could just as easily produce $p(B|A)$ in the last line above by moving the unconditional probabilities to the left-hand side in the last equality.

How is this useful? As an example, hypothetically assume that 2% of the population of the United States are members of some extremist **M**ilitia group ($p(M) = 0.02$), a fact that some members might attempt to hide and therefore not readily admit to an interviewer. A survey is 95% accurate on positive **C**lassification, $p(C|M) = 0.95$, and the unconditional probability of classification (i.e., regardless of actual militia status) is given by $p(C) = 0.05$.* Using Bayes' law, we can now derive the probability that someone positively classified by the survey as being a militia member really *is* a militia member:

$$p(M|C) = \frac{p(M)}{p(C)}p(C|M) = \frac{0.02}{0.05}(0.95) = 0.38. \tag{1.4}$$

The startling result is that although the probability of correctly classifying an individual as a militia member given they really are a militia member

* To illustrate how $p(C)$ is really the normalizing constant obtained by accumulating over all possible events, we will stipulate the additional knowledge that the survey is 97% accurate on negative classification ($P(C^c|M^c) = 0.97$). The unconditional probability of classifying a respondent as a militia member results from accumulation of the probability across the sample space of survey events using the Total Probability Law: $P(C) = P(C \cap M) + P(C \cap M^c) = P(C|M)P(M) + [1 - P(C^c|M^c)]P(M^c) = (0.95)(0.02) + (0.03)(0.98) \cong 0.05$.

is 0.95, the probability that an individual really is a militia member given that they are positively classified is only 0.38.

The highlighted difference between the order of conditional probability is often substantively important in a policy or business context. Consider the problem of designing a home pregnancy test. Given that there exists a fundamental business tradeoff between the reliability of the test and the cost to consumers, no commercially viable product will have perfect or near-perfect test results. In designing the chemistry and packaging of the test, designers will necessarily have to compromise between the probability of **PR**egnancy given positive **T**est results, $p(PR|T)$, and the probability of positive test results given pregnancy, $p(T|PR)$. Which one is more important? Clearly, it is better to maximize $p(T|PR)$ at the expense of $p(PR|T)$, as long as the reduction in the latter is reasonable: it is preferable to give a higher number of false positives, sending women to consult their physician to take a more sensitive test, than to fail to notify many pregnant women. This reduces the possibility that a woman who does not realize that she is pregnant might continue unhealthy practices such as smoking, drinking, and maintaining a poor diet. Similarly, from the perspective of general public health, it is better to have preliminary tests for deadly contagious diseases designed to be similarly conservative with respect to false positives.

1.4.1 Example: Monty Hall

The well-known Monty Hall problem can be analyzed using Bayes' law. Suppose that you are on the classic game show *Let's Make a Deal* with its personable host Monty Hall, and you are to choose one of three doors, A, B, and C. Behind two of the doors are goats and behind the third door is a new car with each door equally likely to provide the car. Thus, the probabilities of selecting the car for each door at the beginning of the game are simply:

$$p(A) = \frac{1}{3}, \qquad p(B) = \frac{1}{3}, \qquad p(C) = \frac{1}{3}.$$

After you have picked a door, say A, before showing you what is behind that door Monty opens another door, say B, revealing a goat. At this point, Monty gives you the opportunity to switch doors from A to C if you want

to. What should you do? The psychology of this approach is to suggest the idea to contestants that they must have picked the correct door and Monty is now trying induce a change. A naive interpretation is that you should be indifferent to switching due to a perceived probability of 0.5 of getting the car with either door since there are two doors left. To see that this is false, recall that Monty is not a benign player in this game. He is deliberately trying to deny you the car. Therefore consider his probability of opening door B. Once you have picked door A, success is clearly conditional on what door of the three possibilities actually provides the car since Monty knows this. So we can define the three conditional probabilities:

The probability that Monty opens door B, given the car is behind A:	$p(B_{Monty}\|A) = \frac{1}{2}$
The probability that Monty opens door B, given the car is behind B:	$p(B_{Monty}\|B) = 0$
The probability that Monty opens door B, given the car is behind C:	$p(B_{Monty}\|C) = 1.$

Using the definition of conditional probability, we can derive the following three joint probabilities:

$$p(B_{Monty}, A) = p(B_{Monty}|A)p(A) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$$

$$p(B_{Monty}, B) = p(B_{Monty}|B)p(B) = 0 \times \frac{1}{3} = 0$$

$$p(B_{Monty}, C) = p(B_{Monty}|C)p(C) = 1 \times \frac{1}{3} = \frac{1}{3}.$$

Because there are only three possible events that cover the complete sample space, and these events are nonoverlapping (mutually exclusive), they form a partition of the sample space. Therefore the sum of these three events is the unconditional probability of Monty opening door B using the Total Probability Law:

$$\begin{aligned} p(B_{Monty}) &= p(B_{Monty}, A) + p(B_{Monty}, B) + p(B_{Monty}, C) \\ &= \frac{1}{6} + 0 + \frac{1}{3} = \frac{1}{2}. \end{aligned}$$

Now we can apply Bayes' law to obtain the two probabilities of interest:

$$p(A|B_{Monty}) = \frac{p(A)}{p(B_{Monty})}p(B_{Monty}|A) = \frac{\frac{1}{3}}{\frac{1}{2}} \times \frac{1}{2} = \frac{1}{3}$$
$$p(C|B_{Monty}) = \frac{p(C)}{p(B_{Monty})}p(B_{Monty}|C) = \frac{\frac{1}{3}}{\frac{1}{2}} \times 1 = \frac{2}{3}.$$

Therefore you are twice as likely to win the car if you switch to door C! This example demonstrates that Bayes' law is a fundamental component of probability calculations, and the principle will be shown to be the basis for an inferential system of statistical analysis.

1.5 Bayes' Law and Conditional Inference

To make the discussion more concrete and pertinent, consider a simple example in sociology and crime studies. One quantity of interest to policy-makers is the recidivism rate of prisoners released after serving their sentence. The quantity of interest is the probability of committing an additional crime and returning to prison. Notice that this is very elusive. Not only are there regional, demographic, and individualistic differences, but the aggregate probability is constantly in flux, given entries and exits from the population as well as exogenous factors (such as the changing condition of the economy).

The traditional analytical approach is to assume that there is some fixed value of the recidivism probability and that it can be estimated with a single point. Conversely, the Bayesian perspective is that this is an unknown quantity that is described in probabilistic terms by a *distribution*, giving probabilities across allowable values. Therefore, the model provides a probabilistic interpretation of the unknown recidivism value.

Looking at data from previous periods, we might have some reasonable guess about the distribution of this probability parameter, $p(A)$. This is the *prior distribution* discussed above. It is termed prior because it is selected before incorporating data into the model.

In all parametric statistical inference, a model is proposed and tested in which a data value has some probability of occurring given a specific value of the parameter. This is the case for both Bayesian and traditional approaches, and is just a recognition that the researcher must specify a

data generation model. Call this quantity $p(B|A)$, meaning the probability of generating a data value at B, given that the parameter value is set at A. For example, the probability of observing the event heads, $B = heads$, in a single toss of a fair coin, $A = 0.5$, is $p(B|A) = 0.5$.

Now consider a third quantity, the unconditional probability of generating given a data value: $p(B)$. This is the probability of observing B without regarding, or averaging over, other facts. So it does not provide any relevant information about more or less likely values of the parameter of interest, A. Stated another way, once the data are observed, information that would lead to good guesses at the likely value of A are contained only in probabilistic statements using both A and B. Second, if we could recover this value later, it might be more convenient to ignore $p(B)$ for now and use it later in Bayesian inference to make the conditional probability $p(A|B)$ sum to one (details on this process are discussed in Chapter 3).

Starting with Bayes' law (1.3), we can calculate $p(A|B) = \frac{p(A)}{p(B)}p(B|A)$. If we temporarily ignore $p(B)$, then:

$$p(A|B) \propto p(A)p(B|A), \tag{1.5}$$

where "$\propto$" means "proportional to" (i.e., the *relative* probabilities are preserved). So the final estimated probability of recidivism (in our running example) given some observed behavior, is proportional to prior notions about the distribution of the probability times the parametric model assumed to be generating the new observed data. The conditional probability of interest on the left-hand side of (1.5) is a balance between things we have already seen or believe, $p(A)$, and the new observations, $p(B|A)$. This is an ideal paradigm for inference in the social and behavioral sciences, since it is consentaneously desirable to build models that test theories with newly observed data, but also based on previous research and knowledge. We never start a data analysis project with absolutely no *a priori* notions whatsoever about the state of nature (or at least we shouldn't!).

The story actually gets better. As the size of the data increases, $p(B|A)$ becomes progressively more influential in determining $p(A|B)$. That is, the greater the number of our new observations, the less important are our previous convictions: $p(A)$. Also, if either of the two distributions, $p(A)$ and $p(B|A)$, are widely dispersed relative to the other, then this distribu-

tion will have less of an impact on the final probability statement. This natural weighting suitably reflects relative levels of uncertainty in the two quantities.

Getting back to our running example, suppose in the past that the recidivism rate was centered around 0.2 and never varied outside of [0.15:0.25]. We then notice that the current data shows great variance in individual behavior but are only able to collect the data for a relatively small number of individuals. This would certainly be evidence that the prior recidivism rate should be trusted more than the currently observable data, which may be an atypical manifestation. Conversely, if the previous recidivism rate varied widely, say 0.1 to 0.9, and the data at hand was very tightly compacted around some value, then we would expect to place more credence in the current data in assigning a distribution to the recidivism parameter. Unlike the built-in machinery of Bayesian statistics, there is no explicit process in classical statistics for making such a tradeoff.

The statistical role of the quantities in (1.5) has not yet been identified. The goal of inference is to make claims about unknown quantities using information currently in hand. Suppose that we designate a generic Greek character to denote an unobserved parameter that is the objective of our analysis. As is typical in these endeavors, we will use θ for this purpose. What we usually have available to us is generically (and perhaps a little vaguely) labeled $\mathbf{D}$ for data. Therefore, the objective is obtain a probabilistic statement about θ given $\mathbf{D}$: $p(\theta|\mathbf{D})$.

Inferences in this book, and in the majority of Bayesian and non-Bayesian statistics, are done by first specifying a parametric model for the data generating process. This defines what the data should be expected to look like given a specific probabilistic function conditional on unknown variable values. These are the common probability density functions (continuous data) and probability mass functions (discrete data) that we already know such as normal, binomial, chi-square, etc. Model specifications indicate the probability of seeing specific data values given fixed parameter values: $p(\mathbf{D}|\theta)$.*

* There is a little fudging here as the probability of obtaining any specific data value with continuous forms is zero. Instead, we should be talking about the density value for a given point, but this distinction will not be relevant to us in practice.

Now we can relate these two conditional probabilities using (1.5):

$$p(\theta|\mathbf{D}) \propto p(\theta)p(\mathbf{D}|\theta). \tag{1.6}$$

But there remains one unknown quantity, $p(\theta)$, the unconditional probability of θ. Where can we get this? There is nothing in this particular term that implies a dependency on the data that we have, and the answer to the question is at the core of Bayesian thinking. The expression $p(\theta)$ is a formalized statement of previous knowledge about θ before observing the data. The basic idea is to specify a *prior* distribution for θ that describes what we know in probabilistic terms and therefore overtly specifying both prior information and uncertainty. If we know little, then this should be a vague probabilistic statement and if we know a lot then this should be a very narrow and specific claim.

The right-hand side of (1.6) implies that the *post*-data inference for θ is a compromise between prior information and the information provided by the new data. The left-hand side of (1.6) is called the *posterior* distribution of θ since it provides the updated distribution for θ *after* conditioning on the data.

In Chapter 3, we go into much greater detail about the mathematical and probabilistic setup of Bayesian inference. The purpose of this brief discussion is to highlight the fact that conditional probability underlies the ability to update previous knowledge about the distribution of some unknown quantity. This is precisely in line with the iterative scientific method that postulates theory improvement through repeated specification and testing with data. The Bayesian approach combines a formal structure of rules with the mathematical convenience of probability theory to develop a process that learns from the data. The result is a powerful and elegant tool for scientific progress in many disciplines.

1.6 Historical Comments

Statistics is a relatively new field of scientific endeavor. In fact, for much of its history it was subsumed to various natural sciences as a combination of fosterchild and household maid: unwanted by its natural parents (mathematics and philosophy), yet necessary to clean things up. Beginning with the work of Laplace (1774, 1781), Gauss (1809, 1823, 1855), Legen-

dre (1805), and de Morgan (1837, 1838, 1847), statistics began to emerge as a discipline worthy of study on its own merits. The first renaissance occurred around the turn of the last century due to the monumental efforts of Galton (1869, 1875, 1886, 1892), Fisher (1921, 1925a, 1925b, 1934), Neyman and Pearson (1928a, 1928b, 1933a, 1933b, 1936a, 1936b), Gossett (as Student, 1908a, 1908b), Edgeworth (1892a, 1892b, 1893a, 1893b), and Pearson (1892, 1900, 1907, 1920). Left out of the twin intellectual developments of frequentist inference from Neyman and Pearson and likelihood inference from Fisher (see Chapter 3, Section 3.3 for details), was the Bayesian paradigm. Sir Thomas Bayes' famous (and only) essay was published in 1763, two years after his death (he chose to perish before publishing). This ingenious work, that precipitated a philosophy about how researcher specified models are fit to data, is the subject of this book.

Fisher in particular was hostile to the Bayesian approach and was often highly critical, though not always with substantiated claims: Bayesianism "which like an impenetrable jungle arrests progress towards precision of statistical concepts." (1922, p. 311). Fisher also worked to discredit Bayesianism and inverse probability (Bayesianism with an assumed uniform prior) by pressuring peers and even misquoting other scholars (Zabell 1989). Yet Fisher (1935) develops *fiducial inference*, which is an attempt to apply inverse probability without uniform priors, but this approach fails; Efron (1998, p. 105) calls this "Fisher's biggest blunder." In fact, Lindley (1958) later proved that fiducial inference is consistent *only* when it is made equivalent to Bayesian inference with a uniform prior. The Neyman-Pearson paradigm was equally unkind to the development of Bayesian statistics, albeit on a less vindictive level. If one is willing to subscribe to the idea of an infinite series of samples, then the Bayesian prior is unimportant since the data will overwhelm this prior. Although there are scenarios where this is a very reasonable supposition, generally these are far more difficult to come by in the social and behavioral sciences.

Although Bayesianism had suffered "a nearly lethal blow" from Fisher and Neyman by the 1930s (Zabell 1989), it was far from dead. Scholars such as Jeffreys (1961), Good (1950), Savage (1954, 1962), de Finetti (1972, 1974, 1975), and Lindley (1961, 1965) reactivated interest in Bayesian methods in the middle of the last century in response to observed deficiencies in

classical techniques. Unfortunately many of the specifications developed by these modern Bayesians, while superior in theoretical foundation, led to mathematical forms that were intractable.* Fortunately, this problem has been largely resolved in recent years by a revolution in statistical computing techniques, and it can be argued that this has led to a second renaissance.

Markov chain Monte Carlo (MCMC) techniques solve a lingering problem in Bayesian analysis. Often Bayesian model specifications considered either interesting or realistic produced inference problems that were analytically intractable because they led to high-dimension integral calculations that were impossible to solve analytically. Beginning with the foundational work of Metropolis et al. (1953), Hastings (1970), Peskun (1973), Geman and Geman (1984), and the critical synthesizing essay of Gelfand and Smith (1990), there is now a voluminous literature on Markov chain Monte Carlo. In fact, modern Bayesian statistical practice is intimately and intrinsically tied to stochastic simulation techniques and as a result, these tools are an integral part of this book.

The basic principle behind MCMC techniques is that if an iterative chain of consecutive values, generated computationally, can be set up carefully enough and run long enough, then *empirical* estimations of integral quantities of interest can be obtained from the later chain values. These Markov chains are successive quantities that depend probabilistically only on the value of their immediate predecessor. In general, it is possible to set up a chain to estimate multidimensional probability structures (i.e., desired probability distributions), by starting a Markov chain in the appropriate sample space and letting it run until it settles into the target distribution. Then when it runs for some time confined to this particular distribution, we can collect summary statistics such as means, variances, and quantiles from the simulated values. This idea has revolutionized Bayesian statistics by allowing the empirical estimation of probability distributions that could not be analytically calculated.

Currently the most popular method for generating samples from posterior distributions using Markov chains is the `WinBUGS` program and its Unix-based precursor `BUGS`. This is a pseudo-acronym for *Bayesian infer-*

* This led one observer (Evans 1994) to compare Bayesians to "unmarried marriage guidance counsellors."

ence Using Gibbs Sampling, referring to the most frequently used method for producing Markov chains. In what constitutes a notable contribution to the Bayesian statistical world, the Community Statistical Research Project at the MRC Biostatistics Unit and the Imperial College School of Medicine at St. Mary's, London provide this high-quality software to users free of charge at: (`http://www.mrc-bsu.cam.ac.uk/bugs/`), and have even made available at the same site extensive documentation by Spiegelhalter, Thomas, Best, and Gilks (1996).

1.7 The Scientific Process in Our Social Sciences

This is a book about the scientific process of discovery in the social and behavioral sciences. Data analysis is best practiced as a theory-driven exploration of collected observations with the goal of uncovering important and unknown effects. This is true regardless of academic discipline. Yet some fields of study are considered more rigorously analytical in this pursuit than others.

The process described herein is that of *inference*: making probabilistic assertions about unknown quantities. It is important to remember that "in the case of uncertain inference, however, the very uncertainty of uncertain predictions renders question of their proof or disproof almost meaningless." (Wilkinson 1977). Thus, confusion sometimes arises in the interpretation of the inferential process as a scientific, investigative endeavor.

Are the social and behavioral sciences truly "scientific"? This is a question asked about fields such as sociology, political science, economics, anthropology, and others. It is not a question about whether serious, rigorous, and important work has been done in these endeavors; it is a question about the research process and whether it conforms to the empirico-deductive model that is historically associated with the natural sciences. From a simplistic view, this is an issue of the conformance of research in the social and behavioral sciences to the so-called scientific method. Briefly summarized, the scientific method is characterized by the following steps:

- Observe or consider some phenomenon.
- Develop a theory about the cause(s) of this phenomenon and articulate it in a specific hypothesis.

- Test this hypothesis by developing a model to fit experimentally generated or collected observational data.
- Assess the quality of the fit to the model and modify the theory if necessary, repeating the process.

This is sometimes phrased in terms of "prediction" instead of theory development, but we will use the more general term. If the scientific method as a process were the defining criterion for determining what is scientific and what is not, then it would be easy to classify a large proportion of the research activities in the social and behavioral sciences as scientific. However useful this typology is in teaching children about empirical investigation, it is a poor standard for judging academic work.

Many authors have posited more serviceable definitions. Braithwaite (1953, p. 1) notes:

> The function of a science, in this sense of the word, is to establish general laws covering the behavior of the empirical events or objects with which the science in question is concerned, and thereby to enable us to connect together our knowledge of the separately known events, and to make reliable predictions of events as yet unknown.

The core of this description is the centrality of empirical observation and subsequent accumulation of knowledge. Actually, "science" is the Latin word for knowledge. Legendary psychologist B. F. Skinner (1953, p. 11) once observed that "science is unique in showing a cumulative process." It is clear from the volume and preservation of published research that social and behavioral scientists *are* actively engaged in empirical research and knowledge accumulation (although the quality and permanence of this foundational knowledge might be judged to differ widely by field). So what is it about these academic pursuits that makes them only suspiciously scientific to some? The three defining characteristics about the *process* of scientific investigation are empiricism, objectivity, and control (Singleton and Straight 1988). This is where there is lingering and often legitimate criticism of the social and behavioral sciences as being "unscientific."

The social and behavioral sciences are partially empirical (data-oriented) and partially normative (value-oriented), the latter because societies develop norms about human behavior, and these norms permeate academic thought prior to the research process. For instance, researchers investi-

gating the onset and development of AIDS initially missed the effects of interrelated social factors such as changes in behavioral risk factors, personal denial, and reluctance to seek early medical care on the progress of the disease as a sociological phenomenon (Kaplan et al. 1987). This is partially because academic investigators as well as health professionals made normative assumptions about individual responses to sociological effects. Specifically, researchers investigating human behavior, whether political, economic, sociological, psychological, or otherwise, cannot completely divorce their prior attitudes about some phenomenon of interest the way a physicist or chemist can approach the study of the properties of thorium: atomic number 90, atomic symbol Th, atomic weight 232.0381, electron configuration $[Rn]7^s26d^2$. This criticism is distinct from the question of objectivity; it is a statement that students of human behavior are themselves human.

We are also to some extent driven by the quality and applicability of our tools. Many fields have radically progressed after the introduction of new analytical devices. Therefore, some researchers may have a temporary advantage over others, and may be able to answer more complex questions: "It comes as no particular surprise to discover that a scientist formulates problems in a way which requires for their solution just those techniques in which he himself is especially skilled." (Kaplan 1964). The objective of this book is to "level the pitch" by making an especially useful tool more accessible to those who have thus far been accordingly disadvantaged.

1.7.1 Bayesian Statistics as a Scientific Approach to Social and Behavioral Data Analysis

The standard frequentist interpretation of probability and inference assumes an infinite series of trials, replications, or experiments using the same research design. The "objectivist" paradigm is typically explained and justified through examples like multiple tosses of a coin, repeated measurements of some physical quantity, or samples from some ongoing process like a factory output. This perspective, which comes directly from Neyman and Pearson, is combined with an added Fisherian fixation with p-values in typical inference in the social and behavioral sciences (Gill 1999).

Very few, if any, social scientists would be willing to argue that human behavior fits this objectivist long-run probability model. Ideas like "personal utility," "legislative ideal points," "cultural influence," and "principal-agent goal discrepancy" do not exist as parametrically uniform phenomena in some physically tangible manner. In direct contrast, the Bayesian or "subjectivist" conceptualization of probability is the degree of belief that the individual researcher is willing to personally assign and defend. This is the idea that an individual *personally* assigns a probability measure to some event as an expression of uncertainty about some event that may only be relevant to one observational situation or experiment.

The central idea behind subjectivist probability is the assignment of a prior probability based on what information one currently possesses and under what circumstances one would be willing to place an even wager. Naturally, this probability is updated as new events occur, therefore incorporating serial events in a systematic manner. The core disagreement between the frequentist notion of objective probability and the Bayesian idea of subjective probability is that frequentists see probability measure as a property of the outside world and Bayesians view probability as a personal internalization of observed uncertainty. The key defense of the latter view is the inarguable point that all statistical models are subjective: decisions about variable specifications, significance thresholds, functional forms, and error distributions are completely nonobjective.* In fact, there are instances when Bayesian subjectivism is more "objective" than frequentist objectivism with regard to the impact of irrelevant information and arbitrary decision rules (c.f. Edwards, Lindman, and Savage 1963, p. 239).

Given the existence of subjectivity in all scientific data analysis en-

* As a brief example, consider common discussions of reported analyses in social science journals and books that talk about reported model parameters being "of the wrong sign." What does this statement mean? The author is asserting that the statistical model has produced a regression coefficient that is positive when it was a priori expected to be negative or vice versa. What is this statement in effect? It is a prior statement about knowledge that existed before the model was constructed. Obviously this is a form of the Bayesian prior without being specifically articulated as such.

deavors*, one should prefer the inferential paradigm that gives the most *overt* presentation of model assumptions. This is clearly the Bayesian subjective approach since both prior information and posterior uncertainty are given with specific, clearly stated model assumptions. Conversely, frequentist models are rarely presented with caveats such as "Caution: the scientific conclusions presented here depend on repeated trials that were never performed," or "Warning: prior assumptions made in this model are not discussed or clarified." If there is a single fundamental scientific tenet that underlies the practice and reporting of empirical evidence, it is the idea that all important model characteristics should be provided to the reader. It is clear then which of the two approaches is more "scientific" by this criterion.

These ideas of what sort of inferences social scientists make are certainly not new or novel. There is a rich literature to support the notion that the Bayesian approach is more in conformance with widely accepted scientific norms and practices. Poirer (1988, p. 130) stridently makes this point in the case of prior specifications:

> I believe that subjective prior beliefs should play a *formal* role so that it is easier to investigate their impact on the results of the analysis. Bayesians must live with such honesty whereas those who introduce such beliefs informally need not.

The core of this argument is the idea that if the prior contains information that pertains to the estimation problem, then we are foolish to ignore it simply because it does not neatly fit into some familiar statistical process. This notion is not particularly controversial among statisticians, as observed by Samaniego and Reneau (1994, p. 957):

> If a prior distribution contains "useful" information about an unknown parameter, then the Bayes estimator with respect to that prior will outperform the best frequentist rule. Otherwise, it will not.

A more fundamental advantage to Bayesian statistics is that both prior and posterior parameter estimates are assumed to have a distribution and therefore give a more realistic picture of uncertainty that is also more useful in applied work:

* See Press and Tanur (2001) for a fascinating account of the role of researcher-introduced subjectivity in a number of specific famous scientific breakthroughs, including discoveries by Galileo, Newton, Darwin, Freud, and Einstein.

> With conventional statistics, the only uncertainty admitted to the analysis is sampling uncertainty. The Bayesian approach offers guidance for dealing with the myriad sources of uncertainty faced by applied researchers in real analyses.

Western (1999, p. 20). Lindley (1986, p.7) expresses a more biting statement of preference:

> Every statistician would be a Bayesian if he took the trouble to read the literature thoroughly and was honest enough to admit he might have been wrong.

This book rests on the perspective, sampled above, that the Bayesian approach is not only useful for social and behavioral scientists, but it also provides a more compatible methodology for analyzing data in the manner and form in which it arrives in these disciplines. As we describe in subsequent chapters, Bayesian statistics establishes a rigorous analytical platform with clear assumptions, straightforward interpretations, and sophisticated extensions. For more extended discussions of the advantages of Bayesian analysis over alternatives, see Berger (1986), Dawid (1982), Efron (1986), Good (1976), Jaynes (1976), and Zellner (1985).

1.8 References

Bayes, T. (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London* **53**, 370-418.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Second Edition. New York: Springer-Verlag.

Berger, J. O. (1986). Bayesian Salesmanship. In *Bayesian Inference and Decision Techniques with Applications: Essays in Honor of Bruno de Finetti*, Arnold Zellner (ed.). Amsterdam: North Holland, pp. 473-488.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. New York: John Wiley & Sons.

Billingsley, P. (1986). *Probability and Measure*. New York: John Wiley & Sons.

Braithwaite, R. B. (1953). *Scientific Explanation*. New York: Harper & Brothers.

Casella, G. and George, E. I. (1992). Explaining the Gibbs Sampler. *American Statistician* **46**, 167-174.

Chung, K. L. (1974). *A Course in Probability Theory*. San Diego: Academic Press.

Dawid, A. P. (1982). The Well-Calibrated Bayesian. *Journal of the American Statistical Association* **77**, 605-613.

de Finetti, B. (1972). *Probability, Induction, and Statistics*. New York: John Wiley & Sons.

de Finetti, B. (1974). *Theory of Probability, Volume 1*. New York: John Wiley & Sons.

de Finetti, B. (1975). *Theory of Probability, Volume 2*. New York: John Wiley & Sons.

de Morgan, A. (1837). Review of Laplace's Théorie Analytique des Probabilités. *Dublin Review* **2**, 338-354; **3**, 237-248.

de Morgan, A. (1838). *An Essay on Probabilities and their Application to Life Contingencies and Insurance Offices.* London: Longman, Orme, Brown, Green, & Longmans.

de Morgan, A. (1847). *Formal Logic; or, the Calculus of Inference, Necessary and Probable.* London: Taylor & Walton.

Draper, D., Hodges, J. S., Mallows, C. L., and Pregibon, D. (1993). Exchangeability and Data Analysis. *Journal of the Royal Statistical Society, Series B* **156**, 9-37.

Edgeworth, F. Y. (1892a). Correlated Averages. *Philosophical Magazine*, 5th Series, **34**, 190-204.

Edgeworth, F. Y. (1892b). The Law of Error and Correlated Averages. *Philosophical Magazine*, 5th Series, **34**, 429-438, 518-526.

Edgeworth, F. Y. (1893a). Exercises in the Calculation of Errors. *Philosophical Magazine*, 5th Series, **36**, 98-111.

Edgeworth, F. Y. (1893b). Note on the Calculation of Correlation Between Organs. *Philosophical Magazine*, 5th Series, **36**, 350-351.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian Statistical Inference for Psychological Research. *Psychological Research* **70**, 193-242.

Efron, B. (1986). Why Isn't everyone a Bayesian? *American Statistician* **40**, 1-11.

Efron, B. (1988). R. A. Fisher in the 21st Century. *Statistical Science* **13**, 95-122.

Evans, S. J. W. 1994). Discussion of the Paper by Spiegelhalter, Freedman, and Parmar. *Journal of the Royal Statistical Society, Series A* **157**, 395.

Feller, W. (1990). *An Introduction to Probability Theory and its Applications.* Volume 1. New York: John Wiley & Sons.

Feller, W. (1990). *An Introduction to Probability Theory and its Applications.* Volume 2. New York: John Wiley & Sons.

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transmissions of the Royal Statistical Society A* **222**, 309-368.

Fisher, R. A. (1925a). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Fisher, R. A. (1925b). Theory of Statistical Estimation. *Proceedings of the Cambridge Philosophical Society* **22**. pp. 700-725.

Fisher, R. A. (1934). *The Design of Experiments.* First Edition. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935). The Fiducial Argument in Statistical Inference. *Annals of Eugenics* **6**, 391-398.

Galton, F. (1869). *Heredity Genius: An Inquiry into its Laws and Consequences.* Second Edition. London: Macmillan.

Galton, F. (1875). Statistics by Intercomparison, with Remarks on the Law of Frequency of Error. *Philosophical Magazine*, 4th Series (49), 33-46.

Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. *Journal of the Anthropological Institute* **15**, 246-263.

Galton, F. (1892). *Finger Prints.* London: Macmillan.

Gauss, C. F. (1809). *Theoria Motus Corporum Clestium.* Hamburg: Perthes et Besser.

Gauss, C. F. (1823). *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae.* Göttingen: Königlichen Gesellschaft der Wissenshcaften.

Gauss, C. F. (1855). *Méthode des Moinddres Carrés. Mémoires sur la Combination des Observations.* Translated by J. Bertrand. Paris: Mallet-Bachelier.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* **85**, 398-409.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**,721-741.

Gill, J. (1999). The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly* **52**, 647-674.

Good, I. J. (1950). *Probability and the Weighting of Evidence.* London: Griffin.

Good, I. J. (1976). The Bayesian influence, or How to Sweep Subjectivism Under the Carpet. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science II*, William L. Harper and Clifford A. Hooker (eds.). Dordrecht: D. Reidel, pp. 125-174.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57**, 97-109.

Jaynes, E. T. (1976). Confidence Intervals vs. Bayesian Intervals. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science II*, William L. Harper and Clifford A. Hooker (eds.). Dordrecht: D. Reidel, pp. 175-257.

Jeffreys, H. (1961). *Theory of Probability.* Oxford, England: Oxford University Press.

Kaplan, A. (1964). *Conduct of Inquiry.* San Francisco: Chandler Publishing Company.

Kaplan, H. B., Johnson, R. J., Bailey, C. A., and Simon, W. (1987). The Sociological Study of AIDS: A Critical Review of the Literature and Suggested Research Agenda. *Journal of Health and Social Behavior* **28**, 140-157.

Kendall, M. G. (1949). On Reconciliation of the Theories of Probability. *Biometrika* **36**, 101-116.

Keynes, J. M. (1921). *A Treatise on Probability.* London: MacMillan.

Knuth, D. E. (1973). *The Art of Computer Programming: Volume 1/Fundamental Algorithms.* Reading, MA: Addison-Wesley.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Berlin: Julius Springer.

Laplace, P. S. (1774). Mémoire sur la Probabilité des Causes par le Évènemens. *Mémoires de l'Académie Royale des Sciences Presentés par Divers Savans* **6**, 621-656.

Laplace, P. S. (1781). Mémoire sur la Probabilités. *Mémoires de l'Académie Royale des Sciences de Paris* **1778**, 227-332.

Laplace, P. S. (1814). *Essai Philosophique sur les la Probabilités.* Paris: V^{e} Courcier.

Legendre, A. M. (1805). *Nouvelles Méthodes Pour la Détermination des Orbites des Comètes.* Paris: Courcier.

Leonard, T. and Hsu, J. S. J. (1999). *Bayesian Methods: An Analysis of Statisticians and Interdisciplinary Researchers.* Oxford, England: Oxford University Press.

Lindley, D. V. (1958). Fiducial Distributions and Bayes' Theory. *Journal of the Royal Statistical Society, Series B* **20**, 102-107.

Lindley, D. V. (1961). The Use of Prior Probability Distributions in Statistical Inference and Decision. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press. pp. 453-468.

Lindley. D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint, Parts 1 and 2.* Cambridge, England: Cambridge University Press.

Lindley, D. V. (1986). Comment. *American Statistician* **40**, 6-7.

Lindley, D. V. and Novick, M. R. (1981). The Role of Exchangeability in Inference. *Annals of Statistics* **9**, 45-58.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of State Calculations by Fast Computing Machine. *Journal of Chemical Physics* **21**, 1087-1091.

Neyman, J. and Pearson, E. S. (1928a). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part I *Biometrika* **20A**, 175-240.

Neyman, J. and Pearson, E. S. (1928b). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part II *Biometrika* **20A**, 263-294.

Neyman, J. and Pearson, E. S. (1933a). On the Problem of the Most Efficient Test of Statistical Hypotheses. *Philosophical Transactions of the Royal Statistical Society, Series A* **231**, 289-337.

Neyman, J. and Pearson, E. S. (1933b). The Testing of Statistical Hypotheses in Relation to Probabilities *a priori. Proceedings of the Cambridge Philosophical Society* **24**. pp. 492-510.

Neyman, J. and Pearson, E. S. (1936a). Contributions to the Theory of Testings Statistical Hypotheses. *Statistical Research Memorandum* **1**, 1-37.

Neyman, J. and Pearson, E. S. (1936b). Sufficient Statistics and Uniformly Most Powerful Tests of Statistical Hypotheses. *Statistical Research Memorandum* **1**, 113-137.

Pearson, K. (1892). *The Grammar of Science.* London: Walter Scott.

Pearson, K. (1900). On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such That It Can Reasonably be Supposed to Have Arisen From Random Sampling. *Philosophical Magazine*, 5th Series, **50**, 157-175.

Pearson, K. (1907). On the Influence of Past Experience on Future Expectation. *Philosophical Magazine*, 6th Series, **13**, 365-378.

Pearson, K. (1920). The Fundamental Problem of Practical Statistics. *Biometrika* **13**, 1-16.

Peskun, P. H. (1973). Optimum Monte Carlo Sampling Using Markov Chains. *Biometrika* **60**, 607-612.

Placket, R. L. (1966). Current Trends in Statistical Inference. *Journal of the Royal Statistical Society, Series A* **129**, 249-267.

Poirer, D. J. (1988). Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics. *Journal of Economic Perspectives* **2**, 121-144.

Press, S. J. and Tanur, J. M. (2001). *The Subjectivity of Scientists and the Bayesian Approach.* New York: John Wiley & Sons.

Samaniego, F. J. and Reneau, D. M. (1994). Toward a Reconciliation of the Bayesian and Frequentist Approach to Point Estimation. *Journal of the American Statistical Association* **89**, 947-957.

Savage, L. J. (1954). *The Foundations of Statistics.* New York: Wiley.

Savage, L. J. (1962). *The Foundations of Statistical Inference.* London: Methuen.

Sheynin, O. B. (1977). Laplace's Theory of Errors. *Archive for History of Exact Sciences* **17**, 1-61.

Singleton, R., Jr. and Straight, B. C. (1998). *Approaches to Social Research.* Third Edition. New York: Oxford University Press.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1996). *BUGS 0.5: Bayesian inference Using Gibbs Sampling Manual (version ii).* MRC Biostatistics Unit:
`http://www.mrc-bsu.cam.ac.uk/bugs/documentation/contents.shtml`.

Skinner, B. F. (1953). *Science and Human Behavior.* Toronto: Macmillan.

Stigler, S. M. (1982). Thomas Bayes' Bayesian Inference. *Journal of the Royal Statistical Society, Series A* **145**, 250-258.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.* Cambridge, MA: Harvard University Press.

Student, A. (1908a). On the Probable Error of a Mean. *Biometrika* **6**, 1.

Student, A. (1908b). On the Probable Error of a Correlation Coefficient. *Biometrika* **6**, 302.

Western, B. (1999). Bayesian Methods for Sociologists: An Introduction. *Sociological Methods & Research* **28**, 7-34.

Wilkinson, G. N. (1977). On Resolving the Controversy in Statistical Inference. *Journal of the Royal Statistical Society, Series B* **39**, 119-171.

Zabell, S. (1989). R. A. Fisher on the History of Inverse Probability. *Statistical Science* **4**, 247-263.

Zellner, A. (1985). Bayesian Econometrics. *Econometrica* **53**, 253-269.

1.9 A Note on the Exercises

At the end of each chapter is a set of exercises of varying difficulty. Each exercise is assigned a difficulty code. This scheme, borrowed from Knuth (1973), gives the level of effort required on a logarithmic scale according to the following categories:

01-09 Trivial

10-19 Simple

20-29 Moderate effort required

30-40 Difficult, significant effort required

40-49 Term project

50 Graduate thesis.

Some of these exercises are also assigned a code (C) indicating whether or not the use of a statistical computing package is required. The actual difficulty levels have been produced by the following Bayesian process: my first guess as to the work involved (the prior), updated by comments and complaints from graduate students at the University of Florida and at the ICPSR Summer Program at the University of Michigan (the data). This is not a perfect process and too much weight may have been placed on the assigned prior since I had dictatorial control.

Consider the following *simple* exercises:

1.1. [**01**] *Who was Bayes?*

1.2. [**05**] *Prove Bayes Theorem.*

1.3. [**04**] *Restate the three general steps of Bayesian inference from page 3 in your own words.*

1.4. [**09**] *Restate Bayes' law when the two events are independent. How do you interpret this?*

1.5. [**10**] *Suppose* $f(\hat{\theta}|\mathbf{X})$ *is the posterior distribution of* θ *given the data* $\mathbf{X}$*. Describe the shape of this distribution when the mode,* $\max_{\theta} f(\hat{\theta}|\mathbf{X})$*, is equal to the mean,* $\int_{\theta} \theta f(\hat{\theta}|\mathbf{X}) d\theta$*.*

1.6. [**C10**] *Run the Gibbs sampling function given below in* `R`*. What effect do you seen in varying the B parameter? What is the effect of producing 5,000 sampled values instead of 500?*

1.10 Computational Addendum: Simple Gibbs Sampling in `R`

As a means of continuing the discussion about conditional probability and covering some basic principles of the `R` language, this addendum introduces the Gibbs sampler (a Markov chain Monte Carlo technique that will be

shown in later chapters to be quite important). The idea behind a Gibbs sampler is to get a marginal distribution for each variable by iteratively conditioning on interim values of the others in a continuing cycle until samples from this process empirically approximate the desired marginal distribution. There will be much more on this topic in Chapter 9 and elsewhere, but here we will implement a trivial but instructive example.

As suggested by Example 2 in Casella and George (1995), suppose that we have two conditional distributions:

$$f(x|y) \propto y\text{exp}[-yx], \quad f(y|x) \propto x\text{exp}[-xy], \quad 0 < x, y < B < \infty. \quad (1.7)$$

These conditional distributions are both exponential probability density functions (see Chapter 3, **Reference Addendum** for details). The upper bound, B, is important since without it there is no finite joint density and the Gibbs sampler will not work. It is possible, but not particularly pleasant, to perform the correct integration steps to obtain the desired marginal distributions: $f(x)$ and $f(y)$. Instead we will let the Gibbs sampler do the work.

The Gibbs sampler is "a transition kernel defined by full conditional distributions" that allows us to run a Markov chain that eventually settles into the desired limiting distribution that characterizes the marginals. In plainer language, it is an iterative process that cycles through conditional distributions until it reaches a stable status whereby future samples characterize the desired distributions. For two parameters, x and y, this involves a starting point, $[x_0, y_0]$, and the cycles defined by drawing random values from the conditionals according to:

$$\begin{array}{ll} x_1 \sim f(x|y_0), & y_1 \sim f(y|x_1) \\ x_2 \sim f(x|y_1), & y_2 \sim f(y|x_2) \\ x_3 \sim f(x|y_2), & y_3 \sim f(y|x_3) \\ \vdots & \vdots \\ \vdots & \vdots \\ x_m \sim f(x|y_{m-1}), & y_m \sim f(y|x_m). \end{array}$$

If we are successful, then after some reasonable period the values x_j, y_j are safely assumed to be empirical samples from the correct marginal dis-

tribution. There are many theoretical and practical concerns that we are ignoring here, and the immediate objective here is to give a rough overview.

The following R function performs the Gibbs sampler for this problem by the following algorithm:

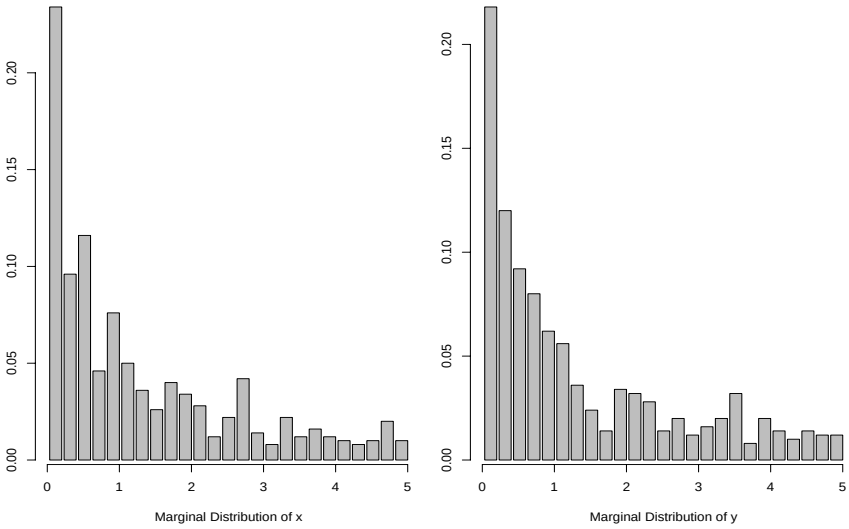
- Set the initial values: $B = 5$, $k = 15$, $m = 500$, and the set of accepted values (x, y) as an empty object. B is the parameter that ensures that the joint distribution is finite, and m is the desired number of total values for x and y. The variable k is the length of the subchains after the starting point, and we will run m separate chains of length k and take the last value as recommended by Gelfand and Smith (1990).
- Run $m = 500$ chains of length $k + 1 = 16$ where the extra value is the uniformly distributed starting point (uniform in $[0 : B]$). Use only sampled exponential values that are less than B and repeat the sampling procedure until such an acceptable value is sampled.
- Add the last value from the x and y series, x_{16} and y_{16} to the list of chain values until 500 of each are obtained.
- Describe the marginal distributions of x and y with these empirical values.

This leads to the following R code, which can be retyped verbatim to replicate this example:

```
B <- 5; k <- 15; m <- 500; x <- NULL; y <- NULL
while (length(x) < m)  {
    x.val <- c(runif(1,0,B),rep((B+1),length=k))
    y.val <- c(runif(1,0,B),rep((B+1),length=k))
    for (j in 2:(k+1))  {
        while(x.val[j] > B) x.val[j] <- rexp(1,y.val[j-1])
        while(y.val[j] > B) y.val[j] <- rexp(1,x.val[j])
    }
    x <- c(x,x.val[(k+1)])
    y <- c(y,y.val[(k+1)])
}
```

These samples are summarized by histograms of the empirical results for x and y in Figure 1.1. It is clear from the figure that the marginal distributions are exponentially distributed. We can recover parameters by using the empirical draws to calculate sample statistics. This part of the

Figure 1.1 Gibbs Sampling Demonstration, Exponentials



MCMC process is actually quite trivial once we are convinced that there has been convergence of the Markov chain. In later chapters we will see this process in a more realistic, and therefore detailed, setting. This example is only intended to give an indication of activities to come and to reinforce the linkage between modern Bayesianism and statistical computing.