That part of Figure 1.1 that is not also part of Figure 1.3 is philosophically outside the range of statistical inference. There is, first of all, the problem of memory failure associated with rectangle 5 and diamond 11. Second, there is the problem that the set of conscious hypotheses or propositions is a small subset of the complete set of propositions. This leads to rectangle 4 and diamond 10, which deal with the elicitation of hypotheses from the enormous file of innate propositions.

Four of the six kinds of specification searches lie within the framework of simple statistical inference: interpretive searches, hypothesis-testing searches, proxy searches and data-selection searches. Simplification searches (postdata) also are a straightforward problem in statistical inference, albeit not in the simple versions. The sixth search—postdata model construction—is either within the framework of statistical inference or not, depending on whether the models that are instigated by the data were conscious or preconscious before the analysis began. If the models were preconscious, inference may usefully proceed as if they were, in fact, conscious, and an inference problem that is necessarily outside the framework of statistical inference can be treated as if it were within.

CHAPTER 2

AN INTRODUCTION TO BAYESIAN INFERENCE

An inference is a logical conclusion drawn from a set of facts. Statistical inference is concerned with drawing conclusions about unobservables $\boldsymbol{\theta}$ from a set of facts, including observed data $\mathbf{x}$ and a conditional probability distribution $f(\mathbf{x}|\boldsymbol{\theta})$, that indicates the probability of various values of $\mathbf{x}$ given various values of $\boldsymbol{\theta}$. Bayesian inference is distinguished from classical inference by its inclusion of a "prior" probability function $f(\boldsymbol{\theta})$ in the set of facts. To a Bayesian there is no sound logical reason why the distribution $f(\mathbf{x}|\boldsymbol{\theta})$ should be regarded to be more of a "fact" than the distribution $f(\boldsymbol{\theta})$. A classicist, however, argues that the distribution $f(\mathbf{x}|\boldsymbol{\theta})$ is an objectively verifiable feature of the world, whereas any distribution $f(\boldsymbol{\theta})$ is purely a figment of someone's imagination. The foundation of the dispute between Bayesians and classicists can thus be found in their definitions of probability, discussed in Section 2.1.

A theme that is developed in this chapter and elsewhere is that data analysis involves three distinct phases. The data $\mathbf{x}$ is first *summarized*, then it is *interpreted*, and lastly *decisions* are made. Summarization and interpretation jointly constitute the process of learning or inference.[1]

[1]Of course, much of the learning activity is aimed explicitly or implicitly at some decision problem, and the sharp distinction between inference and decision is misleading.

$$\text{Data Analysis} \begin{cases} 1.\ \text{Summarization} \\ 2.\ \text{Interpretation} \\ 3.\ \text{Decision} \end{cases} \Big\} \text{Learning}$$

The Bayesian method of data analysis involves each of these three phases. Learning from observations is governed by Bayes' rule

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{x})}$$

which describes how the uncertainty in $\boldsymbol{\theta}$ summarized by the probability function $f(\boldsymbol{\theta})$ is influenced by the data $\mathbf{x}$. Summarization can occur to the extent that $f(\mathbf{x}|\boldsymbol{\theta})$ depends only on some summary of the event $\mathbf{x}$. Interpretation of the evidence $\mathbf{x}$ amounts to changing the uncertainty about $\boldsymbol{\theta}$ from $f(\boldsymbol{\theta})$ to $f(\boldsymbol{\theta}|\mathbf{x})$. Lastly, decisions can be made, given the distribution $f(\boldsymbol{\theta}|\mathbf{x})$.

Classical inference lacks a formal interpretation phase. Strictly speaking, it is only a method of data summarization. Of course, practitioners are interested in learning from data and have built elaborate ad hoc methods of data interpretation. It is hardly surprising that these methods are sometimes in agreement and sometimes greatly at odds with Bayes' rule. Following a brief discussion of Bayes' rule in Section 2.2, Bayesian inferences about a proportion and about a mean are described in Sections 2.3 and 2.4, and the theme of the three phases of data analysis will be elaborated on.

One feature of Bayes' (1763) original essay that brought the greatest scorn was his choice of a prior distribution $f(\boldsymbol{\theta})$ to represent "knowing nothing" (a contradiction in terms?). Opponents of Bayesian inference focus their attacks on the problem of choosing the prior distribution, and Bayesians have responded defensively by trying to find objective subjective priors. My negative attitudes toward the likely fruitfulness of such endeavors are reported in Section 2.5.

2.1 Objective or Subjective Probability?

Classical inference, although based on a seemingly never-ending list of principles, remarkably admits only a single confusing viewpoint, and the principal statistical texts differ mostly in pedagogy and very little in substance. Paradoxically, Bayesian inference, which is based on the single principle of Bayes' rule, admits a basketful of distinctly different viewpoints. The rule straightforwardly describes a way of combining presample (prior) information packaged in a probability distribution with sample information packaged in a likelihood function. Bayesians who accept the rule as their principal commandment find time in their busy missionary schedules to argue vigorously over the Correct Interpretation.

The apparently innocuous rule is simply the conditional probability rule

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

describing the probability of an uncertain event A given the uncertain event B in terms of the probability of B given A, the probability of A and the probability B. What distinguishes Bayesians from non-Bayesians is not their acceptance of the conditional probability rule but rather their willingness to apply it to events A that clearly admit no frequency interpretation. For example, A may be the hypothesis that the gravitational force between two objects decreases with the square of the distance between them. To a Bayesian $P(A)$ summarizes the weight of evidence in favor of A before B is observed, and $P(A|B)$ summarizes the weight of evidence after B is observed. Non-Bayesians argue instead that A is either true or false and that $P(A)$ is appropriately either one or zero, depending on whether A is true or not. Bayes' rule under those circumstances amounts to either $1=1$ or $0=0$.

The distinction between Bayesians and non-Bayesians should thus be understood in terms of the definitions of probability, and it is, therefore, necessary here to discuss the various definitions. The number of conflicting opinions is enormous; for a fuller treatment the reader should consult Barnett (1973, Chap. 3). The viewpoint offered in this book is that probabilists are naturally divided into objectivists, who believe that a probability is usefully regarded as an objective description of physical reality, and subjectivists, who believe that a probability ought to be defined explicitly as a subjective description of man's perception of his surroundings.

THE PROBABILITY AXIOMS

From the standpoint of mathematical theory, probability is a set function that obeys certain axioms. To use the theorems of mathematical probability, it is enough to satisfy yourself that these axioms apply. However, as is discussed subsequently the interpretation of the results of such exercises depends on your understanding of the primitive concept of probability.

Mathematically, probability is described as follows. Let U be a universal or reference set. A function P that associates to every subset $A \subset U$ a real number, $P(A)$, is said to be a *probability measure* on U provided it satisfies the following:

AXIOM 1 For every $A \subset U$, $P(A) \geq 0$

AXIOM 2 $P(U) = 1$

AXIOM 3 If A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$.

These axioms apply in many circumstances in which no one would use the word probability. For example, your arm may contain 10 percent of the weight of your body, but it is unlikely that you would report that the probability of your arm is .1. Objectivists and subjectivists have quite different ideas about the use of the word "probability," and their debate is now to be discussed.

OBJECTIVE PROBABILITY

Objectivists define probability with reference to repetitive phenomena such as dice, roulette, and cards. Although gamblers doubtlessly had some idea how to compute odds in games of chance long before the sixteenth century, it was the Italian mathematician Gerolamo Cardano (1501–1565) who is given credit for the first correct probability calculations. To Cardano, the probability of an event A such as pulling a red card from a deck is simply the ratio of the number of (equally likely) outcomes that lead to the event A divided by the total number of (equally likely) outcomes. This may have been intended only as a formula for calculating probabilities, but deMoivre in 1718 and later Laplace adopted it as a definition, and it is now called the classical definition of probability. As such, it has obvious deficiencies.

To give an example, two flips of a coin can lead to one of three events: a pair of heads, a pair of tails, or a head and a tail. The classical definition might lead us to say that the probability of two heads is one-third. Not so, you "probably" would object; these three events are not equally likely. There are, in fact, four equally likely events: two heads, two tails, a head followed by a tail, and a tail followed by a head. But how are we to know which events are equally likely? And if by equally likely we mean equally probable, have we not circularly presupposed a definition of probability when we defined probability?

Although probability was defined by early writers as a ratio of favorable cases to the total number of cases, the frequency interpretation of probability lurked informally in the background to check the appropriateness of what was meant by "equally likely." This naturally led to a definition of probability in terms of the frequency itself: Let n be the number of trials or experiments (tosses of a coin, rolls of a die) and let m be the number of occurrences of the event A (coin lands heads up, die stops ace up); then we will define the probability of A as

$$P(A) = \frac{\text{number of occurrences of } A}{\text{number of experiments}} = \frac{m}{n}.$$

The ratio m/n, by definition, changes as n changes. If we want to avoid the embarrassment of having our probability assignment to A depend on the number of hypothetical trials, then we must let n grow hypothetically without bound

$$P(A) = \lim_{n \to \infty} \frac{m}{n},$$

that is, we define probability in terms of the *limit* of a *relative frequency*. This, of course, requires that the limit exists, which is, by its definition, impossible to verify. In practice, one always checks his probability assignment by observing the (converging) behavior of m/n in a finite number of trials.

Although it was not until 1837 that Denis Poisson formally defined probability as a limit of a long-run relative frequency, surely gamblers long before the time of Poisson—and before Cardano for that matter—would quote odds based on the relative frequency of occurrence in a limited number of trials. As long as one dealt with repetitive phenomena of a standardized variety—such as in games of chance, actuarial science, genetics, and statistical mechanics—the relative frequency point of view and the classical view equating probability to the ratio of favorable to total cases were adequate. For nonstandardized, nonrepetitive phenomena, however, the frequency definition of probability simply does not usefully apply. A frequentist cannot calculate a nontrivial probability that Andrew Jackson was the eighth President of the United States, or that someone named Andrew Jackson will be the President in the year 2000. To calculate a frequency, we must define the class of relevant experiments. For Andrew Jackson there is (apparently) only one relevant experiment, and the relative frequency is necessarily either one or zero. A frequentist, therefore, would make the following trivial statement. The probability that Andrew Jackson was the eighth President of the United States is either one or zero, one if he was, zero if he was not.

In other cases it may be difficult to define exactly the class of relevant experiments. We often appeal to the vague adjectives "standardized" and "repetitive." All flips of a coin may (intuitively?) be regarded as repetitions of the same experiment. But in a trivial sense, repetition of the same experiment must lead to the same outcome. And a great many things are undeniably different each time we flip a coin. This discussion leads to the conclusion that in order to calculate a relative frequency we must subjectively define the class of events over which to count the frequency. We may all agree on the class of events and in that sense have an "objective" frequency, but that objectivity is something in us and not in nature.

A related problem is that a frequency is a property of a class of events, not a property of individual events. It is quite unclear whether a frequency probability may then be applied to individual events. For example, may we legitimately discuss the probability of getting a head on the next flip of a

coin? There apparently are three positions that frequentists have taken on this issue:

f_1: Probabilities are defined for events that have not occurred. They are not defined for events that have already occurred. We may, therefore, talk about the probability of a head on the next flip of a coin until it is flipped. After that it is either a head or a tail, and no probability applies.

f_2: Probabilities are *not* defined for individual events regardless of whether they have occurred. Probabilities are objective properties of *classes* of events, and they do not apply to individual events.

f_3: Probabilities *are* defined for individual events, both before and after the event occurs under certain circumstances to be discussed subsequently.

The position f_1 makes special reference to the time of occurrence of the event, which is something we may not even know. Imagine that a coin is flipped on a star so distant that it takes 1 hour for light to arrive here. *We* will *see* the flip an hour after it occurs. Is it then the case that for the previous hour we will have the mistaken impression that the probability of a head is one-half, when in fact the probability was not defined since the event had already occurred? More generally, when is an event determined? Precisely when does the probability cease to be defined? When the coin stops vibrating? When it stops rolling but is still vibrating? When it is sailing through the air? When it is resting on the flipping thumb?

Another difficulty with position f_1 is that it can generate probabilities for individual events only in a somewhat circular manner. The frequency is said to apply to an individual event when the experiments determining the hypothetical sequence of events are "standardized and repetitive." If by that statement we have elliptically asserted that all sequences are equally likely, the frequency definition of the probability of an individual event degenerates into the classical definition, with the universe of potential outcomes being all sequences that have a certain frequency. But, as we have already pointed out, that presupposes a definition of probability. It is thus not clear how a probability can be defined for an individual event without first defining probability.

Position f_2 amounts to a negation of the concept of probability as indicating the likelihood of an uncertain event. The usefulness of probability theory is thereby greatly reduced, but there are certainly circumstances in which properties of classes of events are sufficient. It is possible to write insurance, for example, given only the knowledge that 10% of enrollees will suffer some loss. Probabilities applying to particular enrollees are unnecessary. For particular inferences, probabilities applying to particular events seem absolutely essential, however.

The third position f_3 applies the frequency to an individual member of the class when there are no *recognizable subsets* within the class. A subset is recognizable if its frequency is known to differ from the frequency of the complete class of events.

The following illustrates the notion of recognizable subsets. Of all flips of a coin 50% may land heads up, yet it is possible that 75% of those flips that rotate in a primarily north-south direction are heads and only 25% of those flips that rotate in a primarily east-west direction are heads. If this were true and if we knew in which direction a particular coin rotates, then we, clearly, would not say that the probability of a head is one-half, even though the frequency in a large number of trials is one-half. The probability *is* one-half if we do not know which way the coin rotates, that is, if there are no recognizable subsets in the class of all coin flips.

The existence of recognizable subsets is clearly personal, and the position f_3 implies a personal definition of probability. Classical inference built around frequency probability takes either f_1 or f_2. That is, probability statements are made about classes of events not individual events (f_2) or about events that have not occurred (f_1). It is useful here to recall the standard confidence interval statements that students learn to repeat but rarely understand. A 95% confidence interval comes from a class of intervals, 95% of which cover the true value. A particular interval either covers or does not cover the true value, and no probability statement can be made concerning whether it does or does not. This seems to be the position f_2. The position of f_1 is also possible: the probability that an interval covers the true value is .95 until a particular interval is generated. Then the interval either covers, or it does not.

If you are scratching your head in confusion, you have perfectly understood the problem with a frequency definition of probability. I just do not see how a frequentist can make meaningful probability statements at all. He can talk about classes of events and their respective physical, possibly objective, frequencies. But if we reserve the adjective "probability" for set functions that both obey the probability axioms and also indicate the likelihood of uncertain events, a frequentist statement that two times out of ten we will pull a red ball from the urn is no more a probability statement than the statement that 20% of the balls are red or that the red balls make up 20% of the total mass of the balls. Only under certain subjective circumstances can we allow the frequency of .2 to be translated into the statement "The probability of drawing a red ball from the urn is .2."

SUBJECTIVE PROBABILITY

An alternative to the objectivist view that probability is a physical concept such as a limiting relative frequency or a ratio of physically described possibilities is the view first enunciated by James Bernoulli in *Ars Conjectandi* (1713) that probability is a "degree of confidence"—later writers

use degree of "belief"—that an individual attaches to an uncertain event and that this degree depends on his knowledge and can vary from individual to individual. Three important questions arise: (1) Why should they be or why in fact are degrees of belief also probabilities, in the sense of obeying the probability axioms? (2) What is the relationship between degrees of belief and relative frequencies? (3) Are degrees of belief measureable?

Concerning the first question, there are two competing answers. Some statisticians and philosophers assert that a degree of belief is an inherent property of a body of knowledge in the same sense that height is an inherent property of a physical body. As an axiom, they assert that given two uncertain propositions A and B and some body of evidence that relates to the "likelihood" of A and B, one of the following relationships holds: A is more likely, B is more likely, or A and B are equally likely. The word "likely" is not defined by such a statement; only the existence of an ordering is asserted. This is to be compared to the statement that given two individuals it is possible to order them by their height: A is taller than B, B is taller than A, or A and B are equally tall. Height is not thereby defined, but rather is taken as a primitive concept.

Given an individual's ordering of the "likelihood" of uncertain events, we may ask if this ordering is consistent with a probability ordering. That is, do there exist probabilities such that $P(A) < P(B)$ if and only if event A is judged to be less likely than B? Clearly, some orderings of the events rule out probabilities. For example, suppose A is judged more likely than B, B more likely than C, and C more likely than A. It is simply not possible to find numbers $P(A)$, $P(B)$, and $P(C)$ such that $P(A) > P(B)$, $P(B) > P(C)$, and $P(C) > P(A)$. Such an intransitive ordering of events must be ruled out either directly or indirectly.

In fact, several assumptions are necessary for the ordering of uncertain events to be consistent with a probability interpretation. The interested reader is referred to DeGroot (1970, Chap. 6) for a discussion. It is enough here to understand that degrees of belief are asserted to exist and to obey the probability axioms in the sense that certain, apparently compelling, simpler assumptions can be shown to imply the probability axioms. One of these assumptions is briefly discussed: for the probabilities to be unique there must be a reference standard capable of generating all possible probabilities.

Imagine a pointer perfectly balanced on a pin in the middle of a perfect circle as in Figure 2.1. The pointer is to be spun and allowed to come to rest pointing somewhere on the circle. Propositions are arcs such as A and B on the circumference of the circle. The word perfect refers to those requirements that make the degree of belief you hold in any arc dependent

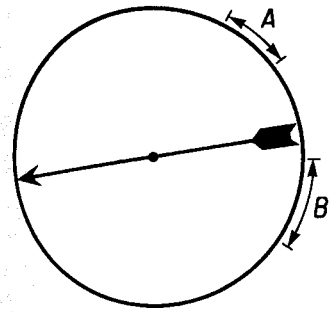

Fig. 2.1 A canonical experiment.

on the length of the arc alone. More than that, a proposition (an arc) A is more probable, equally probable, or less probable than a proposition B if the length of A is greater than, equal to, or shorter than the length of B. If c is the circumference of the circle and $d(A)$ is the length of arc A, we may arbitrarily assign to the degree of belief in A the numerical measure $P(A)$ called the probability of A

$$P(A) = \frac{d(A)}{c} .$$

We now have a reference standard, sometimes called a canonical experiment which can be used to assign particular numbers to the degrees of belief we hold in any propositions. That is, an individual i is said to have degree of belief $P_i(X) = P(A)$ in a proposition X if he regards the proposition X to be equally likely as the arc A in our canonical experiment, where $P(A)$ is the length of arc A divided by the circumference of the circle.

Subjective probability can also be defined in the context of a real or hypothetical problem of decision making under uncertainty. Frank Ramsey (1926) was the first to give a theory of action based on the dual, intertwining notions of judgmental probability and utility. To Ramsey, probability is defined operationally in terms of a person's willingness to act in some decision-making situations in which eventual rewards are uncertain. As an example of such a definition we may say that two uncertain events A and B have the same probability if you are indifferent between winning a dollar if A occurs and winning a dollar if B occurs. For some events there are obvious problems with this definition. Suppose A is the event "Brown will be President of the United States in 1990" and B is the event "the consumer price index will more than quadruple between 1974 and 1990." Although you may regard A and B to be equally likely in some primitive sense, you may prefer to bet on event A since the dollar is likely to be worth more if A is true than if B is true. To give an analogy in

another measurement problem, would you say that individual A is "taller" than B if you would prefer A on your basketball team?

Keeping in mind this difficulty with any decision-based definition of probability, we would like to review here deFinetti's (1937) analysis of betting odds and show in particular that betting odds obey the probability axioms. Suppose you are asked to quote betting odds on a set of uncertain events A, B,..., and accept any wagers others may desire to make about these events. That is, you must assign to each event A a "probability" $P(A)$, thereby indicating a willingness to sell lottery tickets that pay $\$S_a$ if A occurs for the price $\$P(A)S_a$ where S_a are the stakes (positive or negative) to be selected by your opponent. What properties seem desirable for these "probabilities"? Well, you certainly do not want to assign probabilities such that your opponent can select the stakes to guarantee that you will lose regardless of the eventual outcome. This simple *coherence principle* is sufficient to imply the three fundamental axioms of probability:

(a) $1 \geqslant P(A) \geqslant 0$. If your opponent bets only on A and if A occurs, his winnings are $W_1 = S(1-P(A))$, where S may be negative. If A does not occur, he wins $W_2 = -SP(A)$. Coherence requires that $W_1 W_2 \leqslant 0$ for all S. (If $W_1 W_2$ is positive for some S, then W_1 and W_2 have the same sign. If they are both positive, you are a sure loser. If they are both negative, your opponent may change the sign of S to make them both positive.) The condition $W_1 W_2 \leqslant 0$ implies $[1-P(A)]P(A) \geqslant 0$, which implies $0 \leqslant P(A) \leqslant 1$.

(b) $P(U)=1$. The universal set U is certain to occur. Thus your losses on bets about U are, necessarily, $W_u = S_u[1-P(U)]$. By coherence, there must be no S_u such that $W_u < 0$. This implies $P(U)=1$.

(c) If $A \cap B = \phi$, then $P(A \cup B) = P(A) + P(B)$. Suppose you make bets on the events A, B, and $C = A \cup B$. The following events and winnings are possible:

Event	Winnings
$A \cap {\sim}B$	$W_1 = S_a[1-P(A)] - S_b P(B) + S_c[1-P(C)]$
$B \cap {\sim}A$	$W_2 = -S_a P(A) + S_b[1-P(B)] + S_c[1-P(C)]$
${\sim}A \cap {\sim}B$	$W_3 = -S_a P(A) - S_b P(B) - S_c P(C)$

Coherence requires that there be no values of the stakes (S_a, S_b, S_c) such that the winnings (W_1, W_2, W_3) are all positive. If this linear system of equations expressing the winnings as a function of the stakes is invertible, it is possible to specify stakes to make the winnings take on any values whatsoever. To avoid this, the determinant must be zero. Setting the determinant to zero yields

$$0 = \begin{vmatrix} 1-P(A) & -P(B) & 1-P(A \cup B) \\ -P(A) & 1-P(B) & 1-P(A \cup B) \\ -P(A) & -P(B) & -P(A \cup B) \end{vmatrix}$$

$$= -P(A \cup B) + P(A) + P(B).$$

Thus $P(A \cup B) = P(A) + P(B)$.

This treatment of subjective probability, although terribly appealing, has two flaws which have received a certain amount of attention. The first concerns the units in which the stakes S are measured; the second concerns the possibility that the other party to the wager has better information about the event in question. Both are related to the fact that you are not allowed to drop out of the game (when the stakes are too high or when the cards are marked).

Suppose that you are asked to quote odds on the flipping of a coin, with the stakes being a penny. Even odds is the natural choice, and the bet seems fair, if a bit dull. What about stakes of a thousand dollars? Or a million? The nature of the game seems to change as the stakes go up, and what is an acceptable wager for low stakes becomes unacceptable (for most of us) at high stakes. Does this mean that the probabilities change as the stakes go up? No, there is a better explanation. You are interested in happiness, not dollars. If you lose a penny you lose almost the same amount of happiness as you would gain if you were to win a penny. For larger stakes, this is not true. The stakes in terms of happiness are asymmetrical, with even odds of gaining a little and losing a lot. This is exactly the problem discussed previously when probabilities were defined in terms of indifference between lotteries. Expected dollar winnings is not the only thing that matters in choosing lotteries.

We express this graphically in Figure 2.2, which depicts a utility (or happiness) function in terms of dollars. The utility function has the characteristic that continued increments to your wealth provide ever decreasing increments to your happiness. If you do not play the game, you attain happiness level U_0. If you do play, you are equally likely to be at U_w and U_l with U_w only marginally above U_0 but with U_l considerably below U_0. When the stakes are small the discrepancy between $(U_w - U_0)$ and $(U_0 - U_l)$ becomes imperceptible, and the choice between playing and not playing becomes ambiguous.

The point of all this is simply to demonstrate that probabilities can be independent of stakes and also that the stakes can influence the acceptability of a gamble.

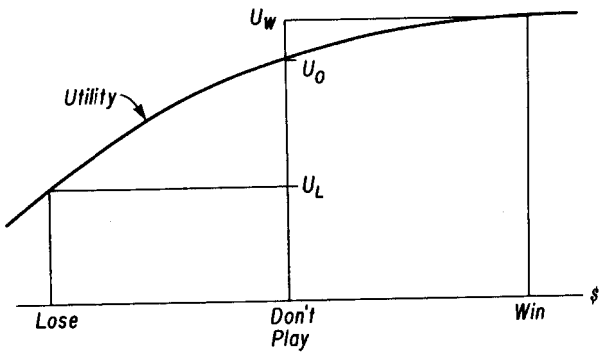

Fig. 2.2 A utility function that discourages gambling.

There is a second problem with this decision-based definition of probability. Although the true probability is naturally thought to be independent of the stakes, the announced probability may not be. The previous discussion implicitly assumes that the second individual has the same opinions as the first and that failure to quote the true probability would result in an advantage to the second player. When the second player has a different probability, the first may want to quote a probability different from the one he truly believes, either to take advantage of the second player or to avoid being taken advantage of himself. The problem is further complicated if the first individual is allowed to drop out of the game. We conclude that, at least for our purposes, it is better not to define probability in the context of some specific decision problem.

This concludes the answer to the first question: why should they be or why in fact are degrees of belief also probabilities? Among the answers are that (1) measures of uncertain knowledge are probabilities for intuitively compelling reasons; or (2) anyone who makes decisions under uncertainty in a "rational" way will act as if he had degrees of belief that obeyed the probability axioms.

Our second question concerns the relationship between degrees of belief and relative frequencies. More generally, what further constraints other than the probability axioms must degrees of belief obey? Consider, for example, the tossing of a coin. It has been generally accepted that a head and a tail are equally likely and are to be assigned an "objective" probability of one-half. The subjective description of probability casts doubt on that statement. After all the central function of a scientific inquiry is to eliminate the uncertainty and to be able to quote longer odds. It is certainly conceivable that a scientist could carefully analyze the coin, landing surface, the size and strength of the flipping thumbs, and the like, and could with confidence conclude that the head is more likely than the

tail on this particular flip. A subjectivist would, therefore, assert that there is no compelling reason why the probability of a head must be one-half.

In fact, subjectivists are divided into two warring factions: the *personalists* and the *necessarists*. Personalists such as deFinetti (1937), Ramsey (1926), and Savage (1954) argue that since knowledge obviously varies from individual to individual, the quantitative measure of knowledge must vary from individual to individual, also. Any individual is advised to constrain his degrees of belief to obey the probability axioms but is otherwise free to assign them as he sees fit (or, more accurately, as is appropriately determined by his knowledge). *Necessarists* such as Jeffreys (1961) and Keynes (1921), on the other hand, argue that a probability is the degree of belief it is *rational* to hold regarding some uncertain proposition, given some other propositions. This is a subtle distinction, if it is a distinction at all. Just what do we mean by "rational"?

Two personalists who had the same joint probability function over all uncertain events would, conditional on the same propositions, have the same (posterior) degrees of belief. Do we mean, then, by rationality that everyone's primitive (preobservation) degrees of belief are identical? This is a conceivable proposition, but it does not seem to have any practical consequences, given that everyone's knowledge (i.e., set of known propositions) is distinctly different. Another possible definition of rationality is errorless computation of degrees of belief. For example, it is "irrational" to believe something to be true merely because you want it to be. But on this point, personalists agree with necessarists; probabilities are descriptions of knowledge measured hypothetically without error, indicating the weight of evidence in favor of uncertain propositions and independent of the desires (or the decision problems) of the individual. A third definition of rational belief is the one that most clearly distinguishes personalists from necessarists. There is a set of propositions that are socially known to be true in the sense that some "reliable" observer includes them in the set he regards to be true. This whole set of propositions is to be used to compute a social or public set of degrees of belief. Probability calculus applies only to these "rational" degrees of belief, because any other degrees of belief are "mistaken." For example, consider the proposition "Andrew Jackson was the eighth President of the United States." A personalist might select an individual whose knowledge about this proposition is incomplete and attempt to measure his degree of belief in the proposition. A necessarist might retort that Jackson either was or was not the eighth President, and the "rational" degree of belief is either one or zero. Any other probability except the correct one is mistaken.

My own feeling is that information transmission among individuals is so imperfect and so poorly understood that it makes little sense to consider public, "rational" degrees of belief. At best, the probability calculus could

apply to personal beliefs, although a substantial part of this book implicitly makes the point that for a variety of reasons real learning is poorly described by this mechanistic mathematical model. At any rate, there are no well-defined "necessary" probabilities.

Although I take the personalist view that probabilities describe personal knowledge, I recognize that there are situations in which many or most people have essentially the same degree of belief in some uncertain proposition. This uniformity of subjective opinion should not be confused with objectivity. The personalist James Dickey has suggested that the words "public" and "private" might informatively replace the words "objective" and "subjective." Public probabilities most often are based on a given relative frequency, in which case a personalist will constrain his probability to be consistent with the publicly known frequency. For example, if it is known that the relative frequency of heads in the next two flips of a coin is exactly one-half, then the personal probabilities of the sequences HH and TT must be zero. Or suppose there is a class of n propositions. The ith proposition, for example, might be "the ith flip of a coin will land heads up." If it is known that exactly f of these propositions are true and if each proposition is equally likely, then the probability of each proposition is, necessarily, f/n. That is, the probability is equal to the relative frequency.

This constraint may be expressed more generally and more formally in the following result, stated in deFinetti (1937). Let the universe consist of m events, $U=\{e_1,e_2,\ldots,e_m\}$, and let the probability function defined on this universe be $P(\{e_k\})$. Suppose that there are n compound events A_1, $A_2,\ldots,A_n$ defined as unions of one or more simple events. Let $P_i=P(A_i)$ and $\pi_j=P$ (exactly j of these events occur). Then it can be shown that

$$\sum_{i=1}^{n}\frac{P_i}{n}=\sum_{j=0}^{n}\frac{j\pi_j}{n}, \tag{2.1}$$

the average probability is equal to the expected relative frequency.

Before this result is proved, note that if the frequency of occurrence j of the n compound events is known to be f with probability one, the right-hand side of this expression becomes just f and the formula can be written

$$\sum_{i=1}^{n}\frac{P_i}{n}=\frac{f}{n}.$$

In words, the average probability is just equal to the relative frequency. Furthermore, if each of the events A_i is equally probable, the formula becomes

$$P_i=\frac{f}{n}.$$

In words, the probability is equal to the relative frequency.

This is an important result that indicates when a personal probability is necessarily equal to a relative frequency. It is a formalization of the notion of recognizable subsets discussed previously. When a class of events is known to have relative frequency f/n and when it is impossible to divide that class into subsets that have greater or lesser probability of occurrence than the class as a whole, we are obligated to adopt the relative frequency as our personal probability.

The proof of the proposition (2.1) is straightforward. Let a_{ik} indicate whether the simple event e_k is included in A_k

$$a_{ik}=\begin{cases}1 \text{ if } e_k\in A_i\\ 0 \text{ otherwise.}\end{cases}$$

Then

$$P_i=\sum_{k=1}^{m}a_{ik}P(e_k).$$

The number of events A_i that occur if the simple event e_k occurs is

$$j(e_k)=\sum_{i=1}^{n}a_{ik}$$

and the expected frequency can be written

$$\begin{aligned}\sum_{j=0}^{n}j\pi_j&=\sum_{k=1}^{m}j(e_k)P(e_k)\\&=\sum_{k=1}^{m}\sum_{i=1}^{n}a_{ik}P(e_k)\\&=\sum_{i=1}^{n}\left(\sum_{k=1}^{m}a_{ik}P(e_k)\right)\\&=\sum_{i=1}^{n}P_i.\end{aligned}$$

This concludes our answer to the second question: what constraints must degrees of belief obey? The answer is, they must be probabilities but are otherwise free. If the relative frequency is known and if each event is

judged to be equally likely, a personal probability will coincidentally equal a frequency.

The third question to be discussed is "are degrees of belief measurable?" Although we hypothesize a complete ordering of uncertain events, we have not indicated how well, if at all, this ordering can be determined. It is useful here to think, again, in terms of heights. We believe that individual A is either taller or shorter than individual B, but we may be unable to identify which one is, in fact, the taller. This is especially true if they are at opposite ends of a football field, but even if we stand them right next to each other, we may have difficulty comparing their heights. Thus the existence of an ordering need not imply the existence of measuring devices sufficiently accurate to disclose the ordering.

The absence of perfect measuring devices does not prevent the use of such concepts as "length," "weight," or "temperature." A designer of a house, for example, at times proceeds as if perfect measurement were possible—he calls this angle a right angle and this length 5 inches, even though no physical angle could be exactly 90° nor any physical length exactly 5 inches. It is essential for him also to consider the effect of likely departures from his design. If the house would fall down unless the angle were 90° to the fifth decimal place, he would alter his design to be more "robust" to departures. We can thus idealize the design of a house into those phases that proceed as if measurement were perfect and those phases that consider the consequences of imperfect measurement.

We take a similar attitude toward probabilities. Although we hypothesize the existence of degrees of belief, it is clearly impossible to measure them without error. It nonetheless makes sense in constructing a theory of inference or a theory of decision making to proceed sometimes as if degrees of belief were measured perfectly. But it is also essential to consider the consequences of measurement error. If our inferential house were to fall down with the slightest discrepancy between the measured and the true degrees of belief, we would surely want to build a different kind of house.

When measurement is very imprecise, it may be argued that we should disdain design altogether and proceed directly to the building of the house. We can expect to learn effective construction in the process, since we will be encouraged to build houses somewhat differently when they fall down. House-building would be taught under such circumstances not by textbooks and lectures but by apprenticeship. In terms of data analysis, it may be argued that the impossibility of measuring degrees of belief makes learning an art, and we should concern ourselves not with designing mathematical models of learning that presuppose perfect measurement of probabilities, but rather with trying different "styles" of learning and

selecting those that turn out best according to some criteria. In principal, I agree with this position, but I think it is useful nonetheless to describe actual learning in formal terms. Mathematical models can be used as teaching devices and perhaps also as guides for improving the learning processes.

We, therefore, consider the inferential implication of certain probabilistic structures suggested by the techniques of data analysis used by economists. We consider the consequences of minor and major changes in the probabilistic structures. And we analyze some forms of measurement error, for example, memory failures.

OBJECTIVISM VERSUS SUBJECTIVISM

We are now in a position to draw the battle line clearly between the objectivist and subjectivist schools. The following list defines functions that obey the probability axioms and which might be called probabilities in the sense of indicating the likelihood of uncertain events:

1. Proportions.
2. Relative frequencies.
3. Degrees of belief.
4. Betting odds.

Representative objectivists and subjectivists were asked to comment on the following pair of statements:

A. The probability of getting a six in the roll of a die is one-sixth.
B. The probability that Andrew Jackson was the eighth president of the United States is one-sixth.

Objectivist. "Statement A is a perfectly good probability statement. Of the six ways that a fair die may land, there is one way favorable to the event 'six'. There is, moreover, no experience with rolls of a die that suggest that "six" occurs more or less often than that. Statement B may reflect someone's betting odds or even be someone's 'degree of belief,' but it is certainly not a probability statement in the same sense as statement A. Jackson either was or was not the eighth President of the United States, a fact we can look up in a book."

Subjectivist. "Both statements may represent someone's degrees of belief or someone's betting odds, and may, therefore, be proper probability statements. Statement A may, on the other hand, be a frequency statement or even just a statement about the proportions of events favorable to the event 'six.' If so, it is no more a probability statement than is the statement

'my legs make up 20% of the weight of my body.' By this I mean that I think the word probability should be restricted to describe the 'likelihood' of uncertain events. A frequency statement indicates only the proportion of a class of propositions that is true. It does not unambiguously indicate the likelihood—the probability—of particular propositions such as 'a six on the next roll of a die.' Of course, there are situations when a frequency statement should be directly translated into a probability statement in my sense. I would think rolling dice would be one of those situations and can well imagine that statement A describes a degree of belief (a probability) as well as a frequency of occurrence."

It is important here to notice that neither the objectivist nor the subjectivist questioned whether the four functions indicated above do, might, or should obey the probability axioms. The subjectivist, in fact, acknowledges that both proportions and frequencies satisfy these requirements. At issue is only which of these functions should be called probability functions in the special sense of indicating the "likelihood" of uncertain events. The objectivist, or more specifically, the frequentist, takes the position that there is a special class of phenomena that lend themselves to probability (frequency) descriptions. He is either uninterested or sees no meaningful mathematical content in statements such as "Andrew Jackson was probably the eighth President." The diametrically opposite position is taken by the subjectivist. He regards the frequency as uninteresting except under those special circumstances when it is also a degree of belief.

In an effort to achieve greater clarity and perhaps a resolution of this conflict we have asked our witnesses the following questions, which as yet have not been answered.

To the *subjectivist*: You have admitted the difficulty of assessing or measuring these abstractions you call degrees of belief. Would you not admit that a person with some confidence can say he holds degree of belief one-sixth in the proposition "a six will occur on the next roll of a die." What person could be so precise about the degree of belief he holds in the proposition "Andrew Jackson was the eighth President" (assuming he was uncertain about it)? In particular, can you not imagine that a person who thinks this proposition is unlikely might *only* be able to say that he holds degree of belief in it less than one-half? Is the frequency/nonfrequency distinction that the objectivist draws thus not also useful to you in thinking about the precision with which degrees of belief may be determined?

To the *objectivist*: The subjectivist has admitted the difficulty of assigning a particular number to many degrees of belief. Does your lack of interest in degrees of belief derive from the apparent difficulty of translating statements such as "Jackson was unlikely to have been the eighth President" into a precise number? Do you not also recognize that any frequency statement makes a number of implicit assumptions concerning the replication of an event? Is it not true that these assumptions are at best approximate and that a frequency statement is thus necessarily somewhat imprecise? And if so, is not your distinction between frequencies and degrees of belief based on a distinction in degree and not in kind?

2.2 Bayes' Rule

The probabilistic rule that plays the central role in Bayesian inference is the conditional probability axiom or "Bayes' rule"

$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

$$=\frac{P(B|A)P(A)}{P(B)}.$$

From the personalist point of view, it indicates that the degree of belief in proposition A given proposition B is equal to the joint probability of A and B divided by the probability of B. It is sometimes called the rule of inverse probability, since it describes how a conditional probability, B given A, can be turned into or inverted into a conditional probability, A given B.

The rule can be written in terms of odds ratios as

$$\frac{P(A|B)}{P(\sim A|B)}=\frac{P(B|A)}{P(B|\sim A)}\frac{P(A)}{P(\sim A)},$$

indicating the posterior odds ratio (given B) equal to the relative likelihood of B under the two hypotheses times the prior odds ratio. The evidential content of B can be completely summarized in terms of the relative likelihood $P(B|A)/P(B|\sim A)$.

The principle of coherence can be used to derive the conditional probability rule (deFinetti, 1937). The first individual establishes his betting odds by announcing a willingness to sell for $\$P(A)S$ lottery tickets which pay $\$S$ in the event that A occurs, where S is the stakes selected by the second individual. The conditional probability $P(A|B)$ should be taken to indicate a willingness to wager on A if B occurs; otherwise, the bet is called off. For ease of notation, we indicate $P(A|B)$, $P(A\cap B)$, and $P(B)$ by p, p', and p'', with the corresponding stakes S, S', and S'' to be chosen by the second individual.

If A is a subset of B, three distinctly different events can occur, and the winnings of the second individual are

$$\begin{aligned} A\cap B \quad & W_1=S'(1-p')+S''(1-p'')+S(1-p)\\ \sim A\cap B \quad & W_2=-S'p' \quad +S''(1-p'')-Sp\\ \sim B \quad & W_3=-S'p' \quad -S''p''. \end{aligned}$$

If this linear system is invertible, the second individual will be able to find values of S, S', and S'' which will make W_1, W_2, and W_3 take on any values whatsoever; in particular, he will be able to inflict arbitrarily large losses on the first individual regardless of the eventual outcome. Thus, coherence requires that p, p', and p'' be selected so that the system is not invertible, that is, so that the determinant $p'-pp''$ is zero. But this is just the conditional probability axiom.

A complete theory of learning is implied by Bayes' rule. A "primitive" joint distribution indicating the personal probability of every event is updated by conditioning on observed events. Learning amounts to merely selecting the appropriate conditional probability. It should be obvious that it is impossible to construct the required joint distribution consciously. Although Bayes' rule may be used unconsciously, it seems unlikely that learning proceeds unconsciously strictly according to Bayes' rule. Three features of this book make overt reference to this author's lack of complete belief in the rule. In Chapter 10 we analyze some simple models of memory failure. In Chapter 9 we discuss "concept formation." In many chapters we report sensitivity analyses which are intended to identify the probabilistic assumptions that crucially determine the nature of the inferences that may be made from a given body of data. In so doing we implicitly admit that no probabilistic assumptions can be made with complete confidence.

2.3 Inference About a Proportion

Consider a population that, of its elements, has a proportion p that possess a given attribute. For example, the population might consist of United States residents of voting age, and p might be the proportion who are currently registered voters. Suppose that the proportion p is not known with certainty, although there is some more or less vague information concerning p. It seems unlikely, for example, that as many as 90% or as few as 10% of the eligible voters are, in fact, registered. Suppose, finally, that n members of the population are asked sequentially if they have the given attribute. What information does the sequence of answers to this question contain concerning the unknown proportion p? What if, for example, no one whom we asked was a registered voter?

The meaning of this sequence of answers depends first of all on how we found the members of the population that we questioned. If we had decided only to ask convicted murderers, it would not be surprising to find no one registered, and the fact that we received such a set of answers would have little impact on our opinions about the proportion p who are registered in the whole population.

Some definitions are now in order. The answers to our query given by the selected population members are called a *sample*. The process by which members of the population are selected for questioning (or sampling) is called the *sampling process*. The set of possible sequences of answers is called the *sample space*. The probability function defined on the sample space indicating the probability of each possible sample is called a *sampling distribution*. Thus the foregoing paragraph indicates that the information content of a particular sample—the extent to which it influences our opinions about p—depends on the sampling process.

The meaning of a sample also depends on prior information about p. Interpretation of a particular sample is quite different if before sampling we thought p was almost exactly .9 than if we thought p was almost exactly .1. We proceed as if our prior opinions about p could be put into a precise distribution, with density, say, $f_1(p)$, indicating that the degree of belief that we hold in the uncertain proposition $a<p<b$ is $\int_a^b f_1(p)dp$. We do, however, want to analyze the extent to which the interpretation of a sample depends on minor changes in the prior density function $f_1(p)$, since, of course, a prior distribution cannot be specified unambiguously.

A word about assumptions is in order. Anyone who insists that personal probabilities be assessed precisely cannot perform statistical inference for the same reason that no spaceship would ever have reached the moon if measurement of lengths had to be perfectly precise before construction could commence. To say something is 6.5 centimeters long is only to say that for the purposes at hand we may proceed *as if* it were. An assumption is not, therefore, a statement of unquestioned truth. It is a tentative statement on which initial action can usefully be based.

We make assumptions about sampling distributions and prior distributions not because they could possibly represent accurately anyone's degrees of belief but rather because they seem sufficiently representative of a class of interesting opinions that they may be used as a useful starting point for an analysis. A statistical analysis is most emphatically a two-way street, however. It involves both the mathematical process of inference given assumptions, and also the artful process of challenging and discarding inadequate assumptions. More is said of this in Chapter 9.

SAMPLING DISTRIBUTIONS

Suppose, first, that only two members of the population are to be sampled. Indicating a positive answer, that is, possession of the given attribute, by S (mnemonic for success) and the contrary event by F (mnemonic for failure) there are four possible outcomes: SS, SF, FS, FF. Both, one, or neither individual may have the attribute. The *sample space* is the set of all possible samples

$$\text{Sample space:} \quad \{SS, SF, FS, FF\}.$$

The probability of obtaining one of the samples—one element in the sample space—depends on how members of the population are identified. The probability function defined on the sample space indicating the probability of each sample is called the *sampling distribution.*

Sampling distribution: $P(SS|p), P(SF|p), P(FS|p), P(FF|p)$.

A special sampling distribution applies if "independent random" sampling is the sampling process. This requires the dual assumption that the probability of selecting a member of the population is independent of the other selections and also that each member is equally likely to be selected at each "draw" from the population. If each member is equally likely to be selected at any given draw, we know from Section 2.1 that the probability of a success on the particular trial (draw, experiment) must be equal to the class frequency p, in this case,

$$P(S|p)=p,\ P(F|p)=1-p.$$

Furthermore, if each draw is independent of the others, the sampling distribution becomes

$$\begin{aligned}P(SF|p)&=P(S|p)P(F|p)=p(1-p)\\P(SS|p)&=P(S|p)P(S|p)=pp\\P(PS|p)&=P(F|p)P(S|p)=(1-p)p\\P(FF|p)&=P(F|p)P(F|p)=(1-p)(1-p).\end{aligned}$$

How do we know if the sampling process yields independent random samples? The answer is, we don't. Remember that these probabilities are personal degrees of belief that, necessarily, are difficult to specify precisely. We *can* identify processes that clearly do not generate independent random samples. The sequence SF may be unambiguously more or less likely to occur than an arc of our canonical experiment covering $p(1-p)$ 100% of the circumference of the circle. We will never be able to say that SF is exactly as likely as this arc, any more than we can say that two rulers are exactly the same length. We will, however, be willing to proceed in many circumstances *as if* we were observing an independent random sample. We must be aware, however, that this is a working hypothesis that ought not to be retained too tenaciously.

Some examples of dependence in the sampling scheme are worth discussing. Suppose the population in question consists of a finite number of members, say, N, with proportion $p=R/N$ possessing the attribute. Suppose we draw from this population one member who has the attribute. If we do not return him to the population, the probability of another success is $(R-1)/(N-1)$. Thus the outcome of the second draw is dependent on the outcome of the first draw. The sampling distribution would then be

$$\begin{aligned}P(SF|p)&=\frac{R}{N}\frac{N-R}{N-1} &&=p(1-p)(1-N^{-1})^{-1}\\P(SS|p)&=\frac{R}{N}\frac{R-1}{N-1} &&=p(p-N^{-1})(1-N^{-1})^{-1}\\P(FS|p)&=\frac{N-R}{N}\frac{R}{N-1} &&=(1-p)p(1-N^{-1})^{-1}\\P(FF|p)&=\frac{N-R}{N}\frac{N-R-1}{N-1} &&=(1-p)(1-p-N^{-1})(1-N^{-1})^{-1}.\end{aligned}$$

Note, however, that for N large this sampling distribution is inconsequentially different from the independent random sampling distribution.

Another form of dependent sampling occurs if bunching of successes and failures is likely. Suppose we wanted to know the proportion of families in Boston that have incomes less than \$5,000, and suppose we identified a "random home" as the first element in our sample. Suppose also that, for the second element of the sample, instead of a random choice, we selected the family who lived next door. The fact that poor people tend to live next door to each other almost guarantees that these neighbors will answer the question identically. The resultant sampling distribution would then be

$$\begin{aligned}P(SF|p)&=0\\P(SS|p)&=p\\P(FS|p)&=0\\P(FF|p)&=1-p.\end{aligned}$$

We henceforth assume independent random sampling. A sample of size n will consist of a sequence of Ss and Fs, n in all. The sample space consists of the 2^n different possible samples, and the sampling distribution is

$$\begin{aligned}P(\text{Sample: } SFFSF\cdots, S|p)&=p(1-p)(1-p)p(1-p)\cdots p\\&=p^r(1-p)^{n-r}\end{aligned}\qquad(2.2)$$

where r is the number of successes and n the sample size.

PRIOR AND POSTERIOR DISTRIBUTIONS OF A PROPORTION

The sampling distributions just discussed describe degrees of belief in propositions about samples given that p is known. The proportion p is also uncertain, and degrees of belief in propositions concerning p may be fully represented in a density function $f_1(p)$, thereby indicating that for any $a \leqslant b$, $P(a \leqslant p \leqslant b)=\int_a^b f_1(p)\,dp$.

Given the sampling distribution (2.2), the prior distribution, and Bayes' rule, we may write the posterior distribution as

$$f_2(p)=f(p|\text{sample})=\frac{P(\text{sample}|p)f_1(p)}{P(\text{sample})}$$

$$=\frac{p^r(1-p)^{n-r}f_1(p)}{P(\text{sample})} \tag{2.3}$$

where P indicates a probability function, f a density function, and where we have been mathematically sloppy in not distinguishing the two more carefully.

The part of this function that deals with the outcome of the sample, namely, $p^r(1-p)^{n-r}$, is called the *likelihood function* of p given that the sample resulted in r successes in n trials. The extent to which the posterior distribution depends on the sample is completely determined by the likelihood function. Multiplying the likelihood function by a constant will not alter the posterior distribution, since it has to be normalized to integrate to one. The likelihood function is, therefore, defined only up to a multiplication constant, and we indicate it as

$$L(p|r,n)\propto p^r(1-p)^{n-r}$$

where $\propto$ indicates "proportional to." Observe that *as a function of* p we have

$$L(p|\text{sample})\propto P(\text{sample}|p).$$

The right-hand side is apparently a probability function defined on the sample space for a particular value of p. The likelihood function is a function of p, however; it is computed from the sampling distribution by identifying how the probability of the particular observed sample depends on p. The fact that the observed sample is twice as probable if $p=p_1$ than if $p=p_2$ is taken as evidence that $p=p_1$. Given this sample, we would say that p_1 is twice as likely as p_2.

The fact that both the likelihood function and the sampling distribution can be written with the same expression causes great confusion for the beginning student, and the reader should examine and understand the following example if he is at all confused by this phenomenon.

Example. A sample consisting of one observation has a sampling distribution $P(S|p)=p$, $P(F|p)=1-p$ depicted graphically in Figure 2.3. If a success is observed, the likelihood function is $L(p|S)\propto P(S|p)=p$, also depicted in Figure 2.3. That is, $p=1$ is most likely; $p=1$ is twice as likely as $p=.5$; $p=0$ is impossible. If, however, a failure is observed, the likelihood function is proportional to $(1-p)$, and $p=0$ is most favored.

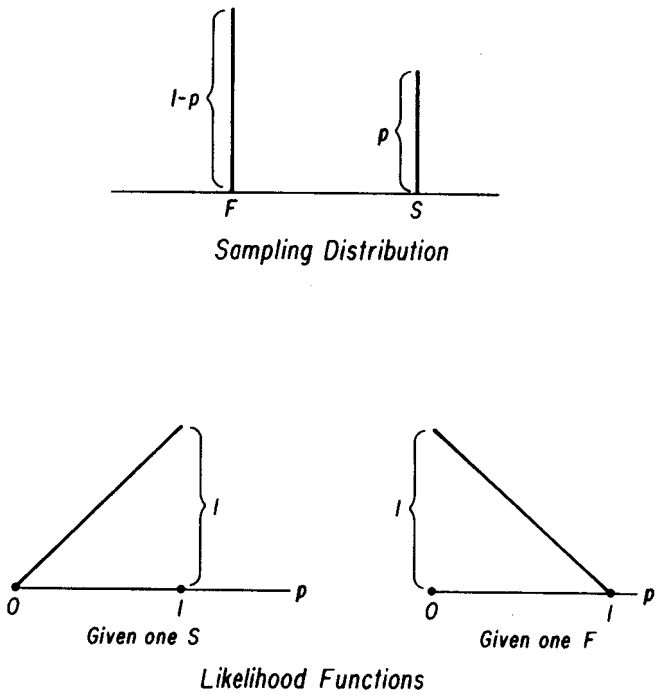

Fig. 2.3 Sampling distributions and likelihood functions.

Now consider again Equation 2.3. Suppose our prior information concerning p is extremely vague in the sense that we have no reason to believe that p is more likely to lie in one interval rather than any other interval of the same length. More specifically, suppose $f_1(p)$ is taken to be rectangular over the interval from zero to one. Then $f_1(p)=1$, and hence by (2.3)

$$f_2(p)=Cp^r(1-p)^{n-r}.$$

By Appendix 2 this p.d.f. is recognized as a beta p.d.f. with parameters $r+1$ and $n+2$, $f_2(p)=f_\beta(p|r+1,n+2)$. More generally, suppose our prior p.d.f. is itself a beta p.d.f. with parameters n_1 and r_1. Then the posterior distribution is also a beta distribution:

THEOREM 2.1. (BINOMIAL SAMPLING AND BETA PRIOR). *If the prior p.d.f. on p is beta with parameters r_1 and n_1, $f_\beta(p|r_1, n_1)$, and a random sample yields n observations, r of which are Ss, then the posterior p.d.f. is*

$$f_2(p)=\frac{1}{C_2}p^{r_2-1}(1-p)^{n_2-r_2-1}=f_\beta(p|r_2,n_2)$$

where

$$n_2 = n_1 + n,$$

and

$$r_2 = r_1 + r.$$

Proof: Substitute

$$f_1(p) = \frac{1}{C_1} p^{r_1 - 1}(1-p)^{n_1 - r_1 - 1}$$

in (2.3) and collect terms.

SOME QUALITATIVE CONCLUSIONS

From (2.3) we see that the shape of the posterior p.d.f. is obtained by multiplying the likelihood function, $L(p|\text{sample})$, by the prior and scaling so that the area under f_2 is unity. The likelihood function is shaped like a beta p.d.f. with parameters $r+1$, $n+2$. Now, by observing graphs of beta p.d.f.'s with different (r,n) combinations, a number of general observations can be made that agree with one's intuition. For example:

(a) As the sample size becomes larger, one's prior distribution and information are "washed out," in the sense that the prior has less and less influence on the posterior.

(b) If one's prior is vague or diffuse, then even with little sample information, the posterior corresponds closely with sample results. (See subsequent discussion.)

(c) If one's prior information is great or his prior distribution "tight," that is, has small variance, say, then reinforcing sample information serves to make the posterior even "tighter."

(d) If one's prior is "tight" and the sample evidence is contradictory, then increasing amounts of such sample information (increasing sample sizes) tends at first to leave the prior relatively unchanged, then may cause the posterior to become more diffuse (representing increasing uncertainty), and finally increasingly large amounts of contradictory information cause a gradual shift to a "tight" posterior position conforming to this sample evidence.

SUMMARIZATION AND INTERPRETATION

We should like for emphasis to point out again that learning can be usefully separated into a summarization phase and an interpretation phase. In this chapter we are discussing inference from a sequence of observable events that can be entirely summarized in terms of two numbers: the

number of successes and the number of trials. More formally, let us define two concepts.

Our data consists of an n-dimensional unknown vector of ones and zeroes where a one indicates success and a zero failure, $(X_1, X_2, \ldots, X_n, X_i = 0, 1)$ distributed as $f(x_1, x_2, \ldots, x_n|p)$ where p is an unknown parameter.

Definition. A *statistic* t is a multivalued real function of the data $X_1, \ldots, X_n$.

Example. In this chapter the data are a sequence of successes and failures. The number of failures is a statistic. Another statistic is the whole sample, that is, an n-dimensional vector of ones and zeroes indicating the trials on which the successes and failures occurred.

Definition. A statistic t is a *sufficient statistic* if the density function of the data can be written as

$$f(x_1, x_2, \ldots, x_n|p) = k(t(x_1, x_2, \ldots, x_n); p)u(\mathbf{x}),$$

where u is a function independent of p and k depends only on $t(\mathbf{x})$, that is, if the likelihood function depends only on t (up to a factor of proportionality).

Example. We have already seen that r and n are sufficient statistics for a binomial p.

A posterior distribution is the product of a likelihood function times a prior distribution. The data can affect the posterior only as they affect the likelihood function, and the likelihood function depends only on a sufficient statistic. Thus a sufficient statistic offers a complete summary of the data evidence. To the extent that we can agree on the process that generates the data, we can agree on the sufficient statistics and thus can agree on the appropriate *summary* of the evidence.

Interpretation of the summarized evidence means changing one's mind, that is, discarding a prior distribution and adopting a posterior distribution. This interpretation obviously depends on the prior distribution in the same way that the summarization depends on the sampling distribution. To illustrate this fact, consider the meaning of 2 successes in 10 trials if p represents one of the following five phenomena:

1. The proportion of times a head occurs in flips of a coin.
2. The proportion of trees in Cambridge that have green leaves on a randomly selected day in 1978.
3. The proportion of Harvard students who have IQs over 150.
4. The proportion of Harvard students who have IQs under 150.
5. The proportion of Martians who weigh more than 150 Marspounds.

Table 2.1

Interpretation of 2 Successes in 10 Trials

Phenomenon	r_1	n_1	95% Prior Interval	P (success on first trial)
Coin lands heads up	50	100	[.419 .581]	1/2
Green leaves	.25	.5	[0, .1] ∪ [.9, 1.0]	1/2
IQ over 150	1	22	[0, .13]	1/22
IQ under 150	21	22	[.867, 1.0]	21/22
Martian weights > 150	0	0	?	?
Phenomenon	r_2	n_2	95% Posterior Interval	P (another success)
Coin lands heads up	52	110	[.417, .579]	52/110
Green leaves	2.25	10.5	[.009, .43]	2.25/10.5
IQ over 150	3	32	[.12, .20]	3/32
IQ under 150	23	32	[.54, .87]	23/32
Martian weights > 150	2	10	[.009, .433]	.2

We have summarized our opinions about these five phenomena in terms of beta distributions with parameters r_1 and n_1 given in Table 2.1. That is:

1. We are almost certain that the proportion of heads is nearly one-half, and our 95% prior interval runs from .419 to .581.
2. Depending on whether the randomly selected day is in winter or in summer, almost all or almost none of the trees will have green leaves. Our 95% interval is, therefore, the union of the intervals [0, .1] and [.9, 1.0].
3. Smart as they are, we do not think many Harvard students have IQs over 150. Our 95% interval runs from 0 to .13.
4. The mirror image of (3).
5. We have very little information about Martians, and have adopted the diffuse prior that is nonintegrable and has no well-defined 95% interval.

The probability of a success on the first trial given p is just p. This is a conditional probability. The marginal probability of a success is just

$$P(\text{success}) = \int P(\text{success}|p) f_\beta(p|r_1, n_1)\, dp$$

$$= \int p f_\beta(p|r_1, n_1)\, dp$$

$$= Ep = \frac{r_1}{n_1}.$$

This may be found in the fourth column of Table 2.1.

The posterior parameters, the posterior 95% interval, and the probability of another success (r_2/n_2) may all be found in the same table. The interpretation of 2 successes in 10 trials is seen to depend on the process that generated the outcomes in the following way.

1. If we are talking about flips of a coin, 2 in 10 is hardly evidence at all. Both the 95% interval and the probability of a success are hardly changed.
2. Two trees in ten with green leaves suggests that it is winter, but there are more green trees than we would have guessed for winter. We thus essentially eliminate the summer branch of our 95% interval [.9, 1.0], but lengthen the winter branch to [.009, .43]. The probability of another success is greatly reduced from .5 to approximately .2.
3. Two in ten students with IQs over 150 is largely corroborative evidence, although it is somewhat more than we expected. We accordingly shorten but adjust rightward our 95% interval from [0, .13] to [.12, .20].
4. Two in ten students with IQs under 150 is startling evidence. [Note this is a different sample from (3), since success is defined differently.] We react to this with confusion by increasing greatly our interval from [.86, 1] to [.54, .86].
5. Where before we were "ignorant" about Martians, we now know a great deal. Notice that Martians and green leaves end up with essentially the same posterior, and arguments over fine adjustments to our definition of ignorance are likely to imply only minor adjustments to the posterior distribution.

EXCHANGEABLE EVENTS AND PREDICTIONS

In formulating Equation (2.2) we have used the independence assumption that if p is known, observation of one outcome does not alter our opinions about the probability of other outcomes. In his paper on probability cited earlier, deFinetti argues that independence is not a phenomenon about which we have very useful intuition. In place of independence, he substitutes the notion of *exchangeability*. A sequence of random events is said to be an *exchangeable sequence* if the probability assigned to particular sequences does not depend on the order of successes and failures, for example, if the sequence SF has the same probability as the sequence FS. The following remarkable theorem is due to deFinetti (1937):

THEOREM 2.2 (INFINITE EXCHANGEABLE SEQUENCE). *Any coherent probability assignment to an infinite exchangeable sequence of binomial events is equivalent to the (marginal) assignment derived from the following joint*

distribution

(*a*) *Conditional on p, the events are drawn independently*

$$P(SSF\cdots FS|p)=p^r(1-p)^{n-r}$$

(*b*) *p has a (unique) (prior) distribution* $f_1(p)$.

The marginal is, of course,

$$P(SSF\cdots FS)=\int_0^1 p^r(1-p)^{n-r}f_1(p)\,dp.$$

Example. A proper proof of this theorem is beyond the scope of this book, but an example can illustrate the ideas. Suppose coins are flipped in such a way that the probability of a head on any flip is one-half. Restricting ourselves to the first two flips and assuming that the sequence is exchangeable, $P(HT)=P(TH)$, the class of proper probability assignments is

Event	Probability	Probability if independent
HH	a	p^2
HT	$\frac{1}{2}-a$	$p(1-p)$
TH	$\frac{1}{2}-a$	$p(1-p)$
TT	a	$(1-p)^2$

for $\frac{1}{2}\geqslant a\geqslant 0$. Theorem 2.2 asserts that there exists a distribution for p such that the expected value of the third column is equal to the second. You may verify that this is true for any distribution such that

$$Ep=\tfrac{1}{2}, Ep^2=a$$

with $Vp=Ep^2-(Ep)^2=a-\frac{1}{4}$.

Note that this cannot be satisfied for $a<\frac{1}{4}$. But for $a<\frac{1}{4}$, we do not have an infinite exchangeable sequence. Suppose $a=0$, for example, then the events HHT, HHH, TTH, and TTT have zero probability, and by the exchangeability assumption, so must HTH, THH, HTT, THT, leaving no event with positive probability. The value $a=0$ is, therefore, not allowable. For two events, two moments of the distribution are determined. For n events, n are determined; hence, the distribution in an infinite sequence is unique.

Suppose now that we have a proper probability assignment on an infinite sequence of exchangeable events. Given that we have observed r successes in the first n trials, the probability of another success is

$$P(S_{n+1}|S_1S_2F_3\cdots F_n)=\frac{P(S_1S_2F_3\cdots F_nS_{n+1})}{P(S_1S_2F_3\cdots F_n)}$$

$$=\frac{\int_0^1 p^{r+1}(1-p)^{n+1-r-1}f_1(p)\,dp}{\int_0^1 p^r(1-p)^{n-r}f_1(p)\,dp}$$

$$=\int_0^1 pf_2(p)\,dp$$

where

$$f_2(p)=\frac{p^r(1-p)^{n-r}f_1(p)}{\int_0^1 p^r(1-p)^{n-r}f_1(p)\,dp}.$$

That is, our prediction of the next outcome is *as if* there were a true proportion p and a prior distribution on p, $f_1(p)$, which we updated according to Bayes' rule to $f_2(p)$ with independent sampling. In fact, both the proportion p and the notion of independence are mathematical fictions (or at least unnecessary accoutrements), since our subjective probability distribution, in fact, applies (need only apply?) to the observable exchangeable sequence of successes and failures. Of course, once we decide that we are observing an infinite exchangeable sequence, it is terribly convenient in formulating our probability assignment to make use of the fictional p and to concentrate our efforts on formulating a prior distribution for p. This sometimes turns back on itself, since in attempting to formulate this prior distribution we properly may ask ourselves questions about the implied distribution on the observable events.

2.4 Inference About a Mean

We have considered in the previous section how inferences may be made about p, the proportion of a given population that possess a given attribute. For attributes that are numerically defined another interesting number is μ, the population mean. For example, μ may be average income, average height, or average IQ. In this section we discuss how inferences may be made about a population mean.

The essential principles of subjectivist inference have been completely established in the previous section. We first select a *sampling* distribution

$P(\text{sample}|\,\mu)$ that describes our predictions about the sample if we knew the population mean μ. For a particular sample this distribution is treated as a function of μ and is called a *likelihood function*. The *posterior distribution* that describes opinions about μ after the sample is observed is computed by multiplying the likelihood function times a *prior distribution* that is selected to summarize our knowledge of μ before we see the sample.

There is one new wrinkle in this section. Interpretation of the evidence about the population mean μ depends on the population variance σ^2. When σ^2 is not known with certainty, we make inferences jointly about both μ and σ^2. That is, we have a two-parameter problem. Although the principles of inference already described do not change, the two-parameter problem is subtly different from the one-parameter problem.

SAMPLING DISTRIBUTIONS

Our first item of business is the construction of a sampling distribution. This is a two-step problem involving, first, the selection of the sample space and, second, the assignment of a probability distribution over that space. In general, we may think of the attribute under study (height, income, etc.) as necessarily lying between minus infinity and plus infinity. If n members of the population are to be sampled, the set of all possible samples is an n-dimensional space

$$\text{sample space} = \{(x_1, x_2, \ldots, x_n) | -\infty < x_i < \infty\}.$$

The very special distribution defined on this space that will take up most of our time is

$$f(x_1, x_2, \ldots, x_n | \mu, \sigma^2) = f_N(x_1 | \mu, \sigma^2) f_N(x_2 | \mu, \sigma^2) \cdots f_N(x_n | \mu, \sigma^2)$$

where $f_N(\cdot | \mu, \sigma^2)$ indicates a normal distribution with mean μ and variance σ^2. This distribution follows from two assumptions:

(1) The population is normal with mean μ and variance σ^2.
(2) The sampling process yields independent random samples.

Both of these assumptions are worth discussing. Let us consider the first. When can we know that the distribution of attributes in a population is normal? Never. Nonetheless, the assumption is a useful starting point. A wide variety of phenomena do have symmetric unimodal distributions that are well approximated by bell-shaped normal curves. In other cases a transformation of the attribute may be approximately normally distributed, for example, the logarithm of income.

Like the assumption of normality, the assumption of independent random sampling can never be asserted to be certainly valid. Various forms of dependence and nonrandomness are likely to be inherent in any sampling process. In some cases departures from independent random sampling will be obvious, but when they are not, assumption (2) makes a useful starting point for the analysis. Again, we emphasize that it is a working hypothesis, not a true proposition.

The sampling distribution can be written as

$$\begin{aligned} f(x_1, x_2, \ldots, x_n | \mu, \sigma^2) &= \prod_{i=1}^{n} f_N(x_i | \mu, \sigma^2) \\ &= \prod_{i=1}^{n} (\sqrt{2\pi}\,\sigma)^{-1} \exp - \frac{(x_i - \mu)^2}{2\sigma^2} \\ &= (\sqrt{2\pi}\,\sigma)^{-n} \exp \frac{-\sum_i (x_i - \mu)^2}{2\sigma^2}. \end{aligned} \tag{2.4}$$

Letting the sample mean be $m = \Sigma x_i / n$, the exponent in (2.4) can be written as

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \mu)^2 &= \sum_{i=1}^{n} [(x_i - m) + (m - \mu)]^2 \\ &= \sum_{i=1}^{n} (x_i - m)^2 + 2(m - \mu) \sum_{i=1}^{n} (x_i - m) + n(m - \mu)^2 \\ &= D^2 + n(m - \mu)^2, \end{aligned} \tag{2.5}$$

where $D^2 \equiv \Sigma_{i=1}^{n} (x_i - m)^2$, and where we made use of the result that $\Sigma(x_i - m) = 0$. Substituting (2.5) into (2.4) we obtain

$$\begin{aligned} f(\mathbf{x} | \mu, \sigma^2) &= \{(2\pi\sigma^2)^{-(n-1)/2} n^{-1/2} e^{-D^2/2\sigma^2}\} \\ &\quad \{(2\pi\sigma^2/n)^{-1/2} e^{-(\mu - m)^2 n / 2\sigma^2}\} \\ &= C f_N(\mu | m, \sigma^2/n) \end{aligned} \tag{2.6}$$

where C stands for the first expression in braces, and where the second expression in braces if $f_N(\mu | m, \sigma^2/n)$. Note that: (1) the expression for C does not depend on μ, and (2) the second expression could equally well have been written as $f_N(m | \mu, \sigma^2/n)$ where μ and m are formally interchanged. We prefer to use the form in (2.6) because of what comes next.

PRIOR TO POSTERIOR ANALYSIS: KNOWN VARIANCE

We have now specified a conditional probability distribution for the sample given the population parameter μ with σ^2 assumed known. A joint distribution over all the uncertain events f (sample, μ) is the product of the

distribution above and a marginal on μ denoted by $f(\mu)$. Symbolically,

$$f(\text{sample}, \mu) = f(\text{sample}|\, \mu) f(\mu)$$

where $f(\mu)$ simply describes our opinions about the population mean before any observations are made. From this joint distribution we may straightforwardly calculate the conditional distribution

$$f(\mu|\text{sample}) = \frac{f(\text{sample}, \mu)}{f(\text{sample})}$$

$$= \frac{f(\text{sample}|\, \mu) f(\mu)}{f(\text{sample})}$$

indicating opinions about μ given any particular sample.

In this context, for obvious reasons, $f(\mu)$ is called a *prior* distribution and $f(\mu|\text{sample})$ is called a *posterior* distribution. Although $f(\text{sample}|\, \mu)$ is a conditional probability distribution, once a particular sample is observed it is treated as a function of μ and called a *likelihood function* denoted by

$$L(\mu; \text{sample}) \propto f(\text{sample}|\, \mu).$$

Note that the *posterior distribution is proportional to the product of the likelihood function and the prior distribution.* The normalizing constant $P(\text{sample})$ is often called a *predictive p.d.f.*, because it summarizes our opinions about the sample before the sample is observed; that is, it predicts the sample.

Making use of (2.6) and Bayes' rule, we may write

$$f_2(\mu) = k f_N(\mu|m, \sigma^2/n) f_1(\mu) \tag{2.7}$$

where k is a proportionality constant and f_2 and f_1 are posterior and prior p.d.f.'s.

Now consider the following two special cases:

CASE 1. Suppose our information concerning μ is extremely vague in the sense that we have no reason to believe that μ lies in any interval rather than any other interval of the same length. Then $f_1(\mu)$ can be taken as a constant in (2.7), and we get

$$f_2(\mu) = f_N(\mu|m, \sigma^2/n) \tag{2.8}$$

since the constant in (2.7) must be selected so that $\int_{-\infty}^{\infty} f_2(\mu)\, d\mu = 1$ and since $\int_{-\infty}^{\infty} f_N(\mu|m, \sigma^2/n)\, d\mu = 1$.

Strictly speaking, a flat prior distribution is not a valid p.d.f. since it does not integrate to one. It therefore violates the principles used to police our opinions about uncertain events. Nonetheless, there are, as we shall see, valid prior p.d.f.'s so nearly flat that for all practical purposes (2.8) is the corresponding posterior distribution.

CASE 2. Suppose our prior is normal with mean m^* and variance v^*. Then by (2.7),

$$f_2(\mu) = k f_N(\mu|m, \sigma^2/n) f_N(\mu|m^*, v^*) \tag{2.9}$$

and substituting the analytical forms for f_N we get, after a lot of elementary algebra, the following result

THEOREM 2.3 (NORMAL SAMPLES AND NORMAL PRIORS). *If the conditional distribution of the vector* $\mathbf{x}$ *given the parameters* μ *and* σ^2 *is normal with mean vector* $E(\mathbf{x}) = \mathbf{1}_n \mu$, *where* $\mathbf{1}_n$ *is a vector of n ones, and variance matrix* $V(\mathbf{x}) = \sigma^2 \mathbf{I}_n$, *and if* μ *is normal with mean* m^* *and variance* v^*, *then the distribution of* μ *given* $\mathbf{x}$ *and* σ^2 *is normal with moments*

$$E(\mu|\mathbf{x}, \sigma^2) = \left(v^{*-1} + (\sigma^2/n)^{-1}\right)^{-1} \left(m^* v^{*-1} + m(\sigma^2/n)^{-1}\right) = m^{**} \tag{2.10}$$

$$V(\mu|\mathbf{x}, \sigma^2) = \left(v^{*-1} + (\sigma^2/n)^{-1}\right)^{-1} = v^{**} \tag{2.11}$$

where m *is the sample mean,* $m = \mathbf{x}'\mathbf{1}_n/n$.

These prior and posterior distributions deserve further comment. (1) Note first that the posterior distribution is in the same family as the prior distribution; that is, both are normal. This is terribly convenient, since the sample evidence can be straightforwardly interpreted in terms of the change in the (two) parameters. If the distribution were to be altered by the sample outcome, there would be rather serious difficulties in discussing the nature of the sample evidence. (2) The posterior mean m^{**} is a weighted average of prior mean and sample mean with weights proportional to v^{*-1} and (n/σ^2), which may be called the precisions of the information sources. Note that for very diffuse priors, that is, for v^* very large, the posterior mean and variance are effectively m and σ^2/n, the sample summaries. This is the situation described in case 1. (3) Suppose that our sample came in two bundles: the first n^* observations yielded a mean of m^*, and the remaining n observations yielded a mean of m. If we started with a uniform prior, after we had seen only the first bundle, we would have opinions about μ that are normal with mean m^* and variance σ^2/n^*. Thus we may think of our prior information with mean m^* and variance v^* as being equivalent to a certain preliminary experiment with

sample size $n^* = \sigma^2/v^*$ and with mean m^*. Note that if formulas (2.10) and (2.11) are defined in terms of $n^* \equiv \sigma^2/v^*$ instead of v^* they become

$$v^{**} = \frac{\sigma^2}{(n^* + n)}$$

$$m^{**} = (n^* + n)^{-1}(n^* m^* + nm),$$

that is, the posterior mean is a weighted average of prior mean m^* and sample mean m with weights proportional to effective prior sample size n^* and sample size n. (4) The posterior variance v^{**} does not depend on the sample mean m. A seemingly desirable property of a posterior distribution is that when the sample and prior distribution are in conflict with neither information source dominating the other, then the posterior distribution should be fairly diffuse. We might also want it to be bimodal. This is not a property of normal sampling with known variance and normal priors, since the posterior distribution is normal with a variance that depends on sample size alone. We may view this as a shortcoming of the normality assumption. (5) The difference between v^* and σ^2 should be clearly understood. The former indicates the uncertainty about μ, and the latter indicates the extent to which a particular observation can wander from μ. There is no reason why these numbers should be the same, or even related.

Sufficiency of the Sample Mean. Given the prior distribution of μ, the posterior distribution of μ depends on the sample outcome $(x_1, \ldots, x_n)$ only through the sample mean

$$m = \sum_{i=1}^{n} x_i/n.$$

We, therefore, say that for prior-to-posterior analysis it is *sufficient* to know *only* the sample mean. This strongly depends on the fact that the sample is known to have been drawn from a normal population with known variance of σ^2 and unknown mean μ.

SUMMARIZATION AND INTERPRETATION

Again, it is useful to emphasize the difference between summarizing and interpreting data evidence. We have assumed a sampling process—the process that generates the data—that admits the sufficient statistic $m = \Sigma x_i/n$. That is, the evidence about μ (when σ^2 and n are known) is completely summarized in the single number m, and, for example, we need not remember the particular value of the first observation x_1 if we remember m. The way that we interpret this summarized evidence depends on our prior information about μ.

To illustrate this, let us form prior distributions for μ when μ is

a. Average age in years of Harvard undergraduates
b. Average age in years of Harvard faculty members
c. Average height in feet of American males
d. Average number of cars per day in thousands in Cambridge during 1978.

We have chosen normal prior distributions for these four averages with means and variances indicated in Table 2.2. The content of this table is

a. We think Harvard undergraduates are most likely to have an average age of 20 years. We are willing to bet at roughly 20:1 odds that the average age is between 19 and 21.
b. Faculty members tend to be older, say, around 50, but most likely have an average age between 40 and 60.
c. The average height of American males is likely to be between 5 and 7 feet.
d. We have very little information about the average number of cars in Cambridge but would guess that it is between 0 and 100,000.

Suppose now we observe a sample with mean $m = 20$ and variance $\sigma^2/n = 5$. The posterior parameters and posterior 95% intervals for each of the four phenomena are also indicated in Table 2.2. The content of these

Table 2.2

Interpreting Evidence From A Normal Sample, $m = 20$, $\sigma^2/n = 5$

	Prior			Posterior		
Phenomenon	m^*	v^*	95% Interval	m^{**}	v^{**}	95% Interval[a]
(a) Age of Harvard undergraduates	20	.25	[19, 21]	20	.24	[19.1, 20.9]
(b) Age of Harvard faculty members	50	25	[40, 60]	25	4.2	[21, 29]
(c) Height of American males	6	.25	[5, 7]	6.7	.24	[5.8, 7.6]
(d) Number of cars in Cambridge	50	625	[0,100]	~20	~5.	[15.5, 24.5]

[a]Approximately.

results is

a. Our opinion about the average age of Harvard undergraduates is supported, though weakly, by the evidence and we narrow our 95% interval to [19.1, 20.9].
b. Observing a sample of Harvard faculty members with average age 20 is astounding. Naive application of the formulas leads to a 95% posterior interval for average age of [21, 29]. It may be more meaningful to question the nature of the process that generated the data.
c. Observing males with average height of 20 feet is even more astounding. Again, we might want to examine the rulers that did the measuring. If we are satisfied with both the sampling process and our prior, we will have a 95% posterior internal of [5.8, 7.6], slightly taller than our prior interval.
d. We began with essentially no information about the number of cars in Cambridge. We now have much better information with our 95% interval reduced from [0,100] to [15.5, 24.5].

We have seen, therefore, that although in all cases the summary of the data is identical, the interpretation of the evidence depends on prior information about the process. We have also seen that naive application of Bayes' rule can lead to some silly results, especially when we have incorrectly described the sampling process. Note also the fact that although the sample supports the prior in Case (a) and contradicts it in Case (c), the posterior variance is the same for both. This is the property of normal sampling with known σ^2 and normal priors that makes us question the wisdom of the normal priors. An apparently desirable property of a learning model is that contradictory evidence induces confusion.

NORMAL SAMPLING WITH UNKNOWN MEAN AND UNKNOWN VARIANCE

When σ^2 is unknown the problem of inference is subtly altered. Beginning with a diffuse prior for μ with σ^2 known, a posterior 95% "credible" interval is given approximately by $m-2\sigma/\sqrt{n} \leqslant \mu \leqslant m+2\sigma/\sqrt{n}$, where m is the sample mean and n is the sample size. The center of this interval m is the sample mean and n is the sample size. The smaller is σ^2, the shorter is our credible interval and consequently the more precise is our opinion about μ. In this sense we may say that m represents the information we have about μ, and that σ^2/n represents the *quality* of that information.

With σ^2 unknown, the *quality* of the information afforded by the sample is necessarily uncertain. As we shall see, the sample contains information about σ^2 we want to use to help "estimate" the quality of the sample

information. We also want to use any available prior information. The interesting feature of this two-parameter problem is that the interpretation of the sample evidence about μ depends on uncertain prior information about σ^2. To the extent that the prior for σ^2 is difficult to specify, the interpretation of the evidence about μ is ambiguous.

For convenience alone, we now would like to replace the process variance σ^2 with a new parameter $h=1/\sigma^2$, called the process precision. We express the prior, likelihood, and posterior in terms of h, keeping in mind that distributions in terms of σ^2 may be easily calculated by a suitable transformation of variables. The likelihood then becomes (from 2.6)

$$L(\mu, h; \text{sample}) = kh^{n/2} e^{-\frac{1}{2}hD^2} e^{-\frac{1}{2}hn(m-\mu)^2}. \tag{2.12}$$

The prior distribution which is most convenient for this likelihood function is a normal-gamma distribution

$$f_1(\mu,h) = f_N\left(\mu|m^*, (hn^*)^{-1}\right) f_\gamma(h|s^{*2}, \nu^*)$$

indicating that conditional on the process precision h, μ has a normal distribution with mean m^* and precision hn^*

$$f_N\left(\mu|m^*,(hn^*)^{-1}\right) = \frac{1}{\sqrt{2\Pi}} (hn^*)^{\frac{1}{2}} e^{-\frac{1}{2}hn^*(\mu-m^*)^2}$$

and h has a gamma distribution

$$f_\gamma(h|s^{*2},\nu^*) = \frac{\left(\frac{1}{2}\nu^* s^{*2}\right)^{\frac{1}{2}\nu^*}}{\left(\frac{1}{2}\nu^*-1\right)!} e^{-\frac{1}{2}\nu^* s^{*2} h} h^{\frac{1}{2}\nu^*-1}.$$

Relevant properties of this gamma distribution are discussed in Appendix 2. Here it may be observed that the marginal on μ is a Student density

$$f(\mu) = \int_0^\infty f_N\left(\mu|m^*,(hn^*)^{-1}\right) f_\gamma(h|s^{*2},\nu^*)\, dh$$
$$= f_S\left(\mu|m^*, \frac{s^{*2}}{n^*}, \nu^*\right)$$

with mean and variance

$$E(\mu) = m^* \text{ and } V(\mu) = \frac{s^{*2}}{n^*}\frac{\nu^*}{\nu^*-2}.$$

We may rewrite the likelihood function as

$$L(\mu,h; \text{sample}) = ke^{-\frac{1}{2}h\nu s^2} h^{\frac{1}{2}\nu} e^{-\frac{1}{2}hn(m-\mu)^2} h^{\frac{1}{2}}$$

where

$$\nu = n - 1$$
$$\nu s^2 = D^2 = \sum (x_i - m)^2.$$

The posterior distribution, as usual, is formed by normalizing the product of the likelihood function and the prior distribution

$$f_2(\mu, h) = kL(\mu, h; \text{sample}) f_1(\mu, h)$$
$$\propto e^{-\frac{1}{2}hns^2} h^{\frac{1}{2}\nu} e^{-\frac{1}{2}hn(m-\mu)^2} h^{\frac{1}{2}}$$
$$h^{\frac{1}{2}} e^{-\frac{1}{2}hn^*(\mu - m^*)^2} e^{-\frac{1}{2}\nu^* s^{*2} h} h^{\frac{1}{2}\nu^* - 1}$$

where irrelevant constants have been subsumed into k. This may be rewritten as

$$f_2(\mu, h) \propto (hn)^{\frac{1}{2}} \exp\left[-\tfrac{1}{2}hn^{**}(\mu - m^{**})^2\right] \exp\left[-\tfrac{1}{2}\nu^{**} s^{**2} h\right] h^{\frac{1}{2}\nu^{**} - 1} \tag{2.12}$$

where

$$\begin{aligned} n^{**} &= n^* + n \\ m^{**} &= (n^* + n)^{-1}(n^* m^* + nm) \\ \nu^{**} &= \nu^* + n \\ \nu^{**} s^{**2} &= \nu^* s^{*2} + n^* m^{*2} + \nu s^2 + nm^2 - n^{**} m^{**2}. \end{aligned} \tag{2.13}$$

By inspection, (2.12) is a normal-gamma distribution. This establishes the following result.

THEOREM 2.4 (NORMAL SAMPLES AND NORMAL-GAMMA PRIORS). *If the conditional distribution of the vector* $\mathbf{x}$ *given* μ *and* σ^2 *is normal with moments* $E(\mathbf{x}|\mu, \sigma^2) = \mathbf{1}_n \mu$ *and* $V(\mathbf{x}|\mu, \sigma^2) = \sigma^2 \mathbf{I}$, *and if* (μ, σ^2) *has a normal-gamma distribution with parameters* $(m^*, n^*, s^{*2}, \nu^*)$, *then* (μ, σ^2) *given* $\mathbf{x}$ *has a normal-gamma distribution with parameters* (2.13).

The first two comments made about sampling with known variance apply here as well. The prior and posterior distributions are in the same four-parameter family of distributions. The posterior mean is a weighted average of prior and sample means. However, the posterior variance now depends on the sample and prior means, since

$$\begin{aligned} \nu^{**} s^{**2} &= \nu^* s^{*2} + \nu s^2 + n^* m^{*2} + nm^2 - n^{**} m^{**2} \\ &= \nu^* s^{*2} + \nu s^2 + \frac{(m - m^*)^2 n^* n}{(n^* + n)}, \end{aligned}$$

and the greater the discrepancy between m and m^*, the larger is s^{**2}, and correspondingly the flatter is the distribution on μ. Thus conflict between the prior and the sample may be so great that the posterior variance of μ exceeds the prior variance. Note also that the variance of μ depends on the prior about σ^2.

2.5 Noninformative Priors

Can a Bayesian compute a posterior distribution for a parameter if he has no prior distribution? Decidedly not; Bayes' rule requires a prior distribution. Well, then, is there a prior distribution that represents ignorance, and can you use Bayesian inference if you are ignorant? Like the issue of original sin, this is a question that remains unresolved, that attracts a fair amount of theological interest, but that seems rather remote from the concerns of the man on the street.

The argument begins chronologically with Bayes' *An Essay Towards Solving a Problem in the Doctrine of Chances*, which suggests by the principle of insufficient reason that ignorance is represented by a probability function that assigns equal probability to all events. Of course, there can be no such probability function, since if mutually exclusive events A and B are assigned equal probability, the event A union B is implicitly assigned twice the probability. Or if a continuous random variable θ is assigned a uniform distribution, the variable $\gamma = \theta^{-1}$ has density function proportional to γ^{-2}. In a situation of "real ignorance" there is insufficient reason to select one event space rather than another, or one parameterization rather than another, and the principle of insufficient reason is apparently insufficient to determine probabilities.[2]

Jeffreys (1961) is the modern-day proponent of the Bayes–LaPlace principle of insufficient reason. He suggests that for a parameter μ defined from $-\infty$ to ∞, one should use the (improper) uniform prior distribution

$$f(\mu)\, d\mu \propto d\mu, \quad -\infty < \mu < \infty.$$

This seems to me to be implying that μ almost certainly is either enormously large positive or enormously large negative, since the ratio of the mass outside any finite interval to the mass inside any finite interval is one. A researcher usually can get away with making such a ridiculous statement in practice, since the data will imply even more strongly that the parameter μ is not enormous.

[2]Are individuals A and B equally tall, merely because there is insufficient reason to regard one to be taller than the other?

This last observation suggests a philosophically sound method of computing a diffuse prior distribution. Assuming that a particular experiment is about to be observed, we can find a prior distribution that will have little impact on the posterior distribution. It is important to understand that this is not a prior intended to represent ignorance. It is defined only in the context of a particular experiment. It is a prior that represents information that will be dominated by the sample information. For example, referring to formula (2.10) that describes the parameters of the posterior distribution for μ, observe that as the sample size increases and the sample information dominates the prior information, the parameters converge to n and σ^2; but if we set v^* to zero and use a uniform prior for μ, the posterior and prior will bare this same relationship to each other regardless of the sample size. Thus a uniform distribution is a prior that is dominated by any sample from a normal population.[3]

In practice, the sample may dominate the prior information and the posterior distribution may be inconsequentially different from a posterior distribution corresponding to an improper noninformative prior. A prior density that is relatively uniform where the likelihood function attains its maximum is likely to imply such a posterior. This is discussed by Savage (1962) under the title, "stable estimation." But for any proper prior, there are samples that do not generate information that dominates the prior, and it is impossible to know in advance of observation whether a diffuse prior is an adequate approximation to the proper prior that is truly representative of your opinions.

Of course, there are difficulties in precisely defining a noninformative prior. But much more important is the fact that the use of improper, noninformative priors can sometimes lead to very undesirable outcomes, many of which are discussed by Lindley (1971). In the hypothesis-testing problem of Chapter 4, noninformative priors imply that all hypotheses with more than the minimal number of parameters should certainly be rejected. A noninformative prior can lead to inadmissible decision rules, as discussed in Chapter 5. Quite pragmatically, however, since we have gone to all this trouble to develop a method that formally incorporates subjective prior information, it seems a shame in the end to use priors that are intended to represent ignorance. Is the Bayesian logic to be used only to rename the likelihood function?

[3]Another way of defining priors in relation to anticipated sample information is due to Jeffreys. He suggests that if a researcher is ignorant about a parameter θ, then his opinions about θ given some evidence $\mathbf{x}$ should be the same whether he regards θ to be the parameter or some one-to-one differentiable transformation of it, $g(\theta)$. A prior that has this invariance property is

$$f(\theta) \propto \left| -E\left(\partial^2 \log f(x|\theta)/\partial\theta^2\right)\right|^{1/2}$$

where the expectation is with respect to the density $f(\mathbf{x}|\theta)$. In words, the prior is proportional to the square root of Fisher's information measure. For a proof and discussion see Zellner (1971) where minimal information priors are also discussed.

One reason statisticians have spent so much time looking for priors to represent ignorance is that in practice it is extremely difficult actually to select a prior. The word "ignorance" is suggestive of a prior that is not only relatively uninformed but also difficult to determine with adequate accuracy. It may make sense in that case to let the data help in measuring the prior by determining a class of dominated priors, all of which imply essentially the same posterior distribution as does the diffuse improper prior. One need only ask himself if this particular sample outcome seems "peculiar." If the answer is no, one might as well use the posterior distribution implied by the diffuse prior, since careful measurement of the prior can lead to only minor adjustment of the posterior.