differences between the probability models that the Bayesians use and the ones that the frequentists use to articulate their very different representations of the discipline of statistics. We emphasize that the likelihood methods advocated here require only the frequentists' probability models, and do not entail Bayesian prior probability distributions for parameters.

CHAPTER 1

The first principle

1.1 Introduction

In this chapter we distinguish between the specific question whose answer we seek and other important statistical questions that are closely related to it. We find the answer to our question in the simplest possible case, where the proper interpretation of statistical evidence is transparent. And we begin to test that answer with respect to intuition, or face-validity; consistency with other aspects of reasoning in the face of uncertainty (specifically, with the way new evidence changes probabilities); and operational consequences. We also examine some of the common examples that have been cited as proof that the answer we advocate is wrong. We observe two general and profound implications of accepting the proposed answer. These suggest that a radical reconstruction of statistical methodology is needed. Finally, to define the concept of statistical evidence more precisely, we illustrate the distinction between degrees of uncertainty, measured by probabilities, and strength of evidence, which is measured by likelihood ratios.

1.2 The law of likelihood

Consider a physician's diagnostic test for the presence or absence of some disease, D. Suppose that experience has shown the test to be a good one, rarely producing misleading results. Specifically, the performance of the test is described by the probabilities shown in Table 1.1. The first row shows that when D is actually present, the test detects it with probability 0.95, giving an erroneous negative result with probability 0.05. The second row shows that when D is absent, the test correctly produces a negative result with probability 0.98, leaving a false positive probability of only 0.02.

Now suppose that a patient, Mr Doe, is given the test. On learning that the result is positive, his physician might draw one of the

Table 1.1 *A physician's diagnostic test for the presence or absence of disease D*

		Test result	
		Positive	Negative
Disease D	Present	0.95	0.05
	Absent	0.02	0.98

following conclusions:

1. Mr Doe probably does not have D.
2. Mr Doe should be treated for D.
3. The test result is evidence that Mr Doe has D.

Which, if any, of these conclusions is appropriate? Can any of them be justified? It is easy to see that under the right circumstances all three might be simultaneously correct.

Consider conclusion 1. It can be restated in terms of the probability that Mr Doe has D, given the positive test, $\Pr(D|+)$; it says that $\Pr(D|+) < \frac{1}{2}$. Whether this is true or not depends in part on the result (+) and the characteristics of the test (Table 1.1). But it also depends on the prior (before the test) probability of the condition, $\Pr(D)$. Bayes's theorem shows that

$$\Pr(D|+) = \frac{\Pr(+|D)\Pr(D)}{\Pr(+|D)\Pr(D) + \Pr(+|\text{not-}D)\Pr(\text{not-}D)}$$
$$= \frac{0.95\Pr(D)}{0.95\Pr(D) + 0.02(1 - \Pr(D))}.$$

If D is a rare disease, so that $\Pr(D)$ is very small, then it will be true that $\Pr(D|+)$ is small and conclusion 1 is correct (as, for example, if $\Pr(D) = 0.001$, so that $\Pr(D|+) = 0.045$). On the other hand, if D were more common – say, with a prior probability of $\Pr(D) = 0.20$ – then $\Pr(D|+)$ would be 0.92, and conclusion 1 would be quite wrong. The validity of conclusion 1 depends critically on the prior probability.

Even if conclusion 1 is correct – say, $\Pr(D|+) = 0.045$ – conclusion 2 might also be correct, and the physician might appropriately decide to treat for D even though it is unlikely that D is present. This might be the case when the treatment is effective if D is present but harmless otherwise, and when failure to treat a patient who actually

has D is disastrous. But conclusion 2 would be wrong under different assumptions about the risks associated with the treatment, about the consequences of failure to treat when D is actually present, etc. It is clear that to evaluate conclusion 2 we need, in addition to the information required to evaluate conclusion 1, to know what are the various possible actions and what are their consequences in the presence of D and in its absence.

But how about conclusion 3? The rule we will consider implies that it is valid, independently of prior probabilities, and without reference to what actions might be available or their consequences: the positive test result is evidence that Mr Doe has the disease. Furthermore the rule provides an objective numerical measure of the strength of that evidence.

We are concerned here with the interpretation of a certain kind of observation as evidence in relation to a certain kind of hypothesis. The observation is of the form $X = x$, where X is a **random variable** and x is one of the possible values of X. We begin with hypotheses which, like the two in the example of Mr Doe's test, imply definite numerical probabilities for the observation. Later we will consider more general hypotheses.

> *Law of likelihood*: If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x)/p_B(x)$, measures the strength of that evidence (Hacking, 1965).

In our example the hypothesis (A) that Mr Doe has disease D implies that a positive test result will occur with probability 0.95, while hypothesis B, that he does not have D, implies that the probability is only 0.02. Thus, according to the law of likelihood, Mr Doe's positive test is evidence supporting A over B, and conclusion 3 is correct.

1.3 Three questions

The physician's three conclusions can be paraphrased as follows:

1. I believe B to be true.
2. I should act as if A were true.
3. This test result (+) is evidence supporting A over B.

These are answers to three generic questions:

1. What do I believe, now that I have this observation?
2. What should I do, now that I have this observation?
3. What does this observation tell me about A versus B? (How should I interpret this observation as evidence regarding A versus B?)

Cox (1958) distinguished between the problem areas represented by questions 2 and 3 and emphasized the importance of the latter:

> Even in problems where a clear-cut decision is the main object, it very often happens that the assessment of losses and prior information is subjective, so that it will help to get clear first the relatively objective matter of what the data say... In some fields, too, it may be argued that one of the main calls for probabilistic statistical methods arises from the need to have agreed rules for assessing strength of evidence.

The third question is the one we want to answer. Although all three are obviously important, we will consider the first two only to clarify the third. It is the third question that is central to the reporting of statistical data in scientific journals. For example, an epidemiologist might investigate the risk of a certain disease among workers exposed to a chemical agent in comparison to the risk among unexposed workers. He produces a data set, and our objective as statisticians is to understand how the data should be presented and interpreted as evidence about the risks. Suppose it has been hypothesized that exposure might be associated with a substantial increase in the risk of the disease. Are these data evidence supporting that hypothesis? If so, how strong is the evidence for, say, a fivefold increase versus no increase? Is this evidence consistent with that found in other studies? If the published report presents clear answers to such questions then it will be helpful to readers who will use this evidence, along with that from other sources, in deciding whether to move for changes in the workplace, whether to do another, larger, study, whether to undertake an investigation to explain how the chemical exposure might lead to the disease, whether to change jobs, etc. The published paper presents the data, along with analyses that make clear its evidential meaning. The readers will then use the evidence to adjust their beliefs and to help them in making decisions.

We will concentrate on hypotheses of a special kind, statistical hypotheses. A **simple statistical hypothesis** is one that completely specifies the probability distribution of an observable random variable. A **composite statistical hypothesis** asserts that the distribution belongs to a specified set of distributions. In our diagnostic example the random variable X represents the outcome of Mr Doe's test, and the two hypotheses about the presence or absence of D imply two simple statistical hypotheses: if D is present then X has the probability distribution given in the first row of Table 1.1, and if D is absent then X has the distribution given in the second row. When the observations are not numerical, as in Mr Doe's test where the outcomes are 'positive' or 'negative', we will usually give them numerical codes such as $1 \equiv$ 'positive' and $0 \equiv$ 'negative'. The random variable will often be vector-valued, i.e. a realization x of X is not a single number, but an ordered set of numbers, as it would be if we observed not only Mr Doe's test result (x_1) but also his blood pressure (x_2) and pulse rate (x_3). Then the observation would be a vector $x = (x_1, x_2, x_3)$.

The reader might have noticed that the law of likelihood, as stated, does not apply to continuous probability distributions. This limitation is not essential, and Exercise 1.1 extends it to continuous distributions. But for now we must see if the law is persuasive in the simple discrete case.

1.4 Towards verification

Why should we accept the law of likelihood? One favorable point is that it seems to be the natural extension, to probabilistic phenomena, of scientists' established form of reasoning in deterministic situations. If A implies that under specified conditions x will be observed, while B implies that under the same conditions something else, not x, will be observed, and if those conditions are created and x is seen, then this observation is evidence supporting A versus B. This is the law of likelihood in the extreme case of $p_A(x) = 1$ and $p_B(x) = 0$. The law simply extends this way of reasoning to say that if x is more probable under hypothesis A than under B, then the occurrence of x is evidence supporting A over B, and the strength of that evidence is determined by how much greater the probability is under A. This seems both objective and fair – the hypothesis that assigned the greater probability to the observation did the better job of predicting what actually happened, so it is better supported by that observation. If the likelihood ratio, $p_A(x)/p_B(x)$, is very large, then hypothesis A did a much better job than B of predicting which value X would take, and the observation $X = x$ is very strong evidence for A versus B.

One crucial test of the law of likelihood is for consistency with the rules of probability theory. There are serious questions about when it is meaningful to speak of the probability that a hypothesis A is true. But there certainly are some situations where hypotheses have probabilities. (For example, if I generate X by drawing balls from one urn or another, and if I choose which urn to draw from by a coin toss, then the hypotheses corresponding to the two urns both have probability 0.5.)

Suppose A and B are hypotheses for which $\Pr(A)/\Pr(B)$ is the probability ratio before X is observed. The elementary rules governing conditional probabilities imply that after $X = x$ is observed, the probability ratio is changed to

$$\frac{\Pr(A|X=x)}{\Pr(B|X=x)} = \frac{\Pr(X=x|A)\Pr(A)}{\Pr(X=x|B)\Pr(B)} = \frac{p_A(x)}{p_B(x)}\frac{\Pr(A)}{\Pr(B)}. \qquad (1.1)$$

This shows that the new evidence, that the observed value of the random variable X is x, changes the probability ratio by the factor $p_A(x)/p_B(x)$, precisely in agreement with the law of likelihood. If we use the law then our interpretations of data as evidence will be consistent with the rules of probability theory; we will never claim that an observation is evidence supporting A over B when the effect of that observation, if A and B had probabilities, would be to reduce the probability of A relative to that of B. Furthermore, the factor $p_A(x)/p_B(x)$, that the law uses to measure the strength of the evidence, is precisely the factor by which the observation $X = x$ would change the probability ratio $\Pr(A)/\Pr(B)$.

It is important to be aware that in asking which is better supported, A or B, we are not assuming that one or the other must be true. On this point, we note that equation (1.1) does not require that the two probabilities, $\Pr(A)$ and $\Pr(B)$, sum to one.

Another crucial test of the law of likelihood is operational – does it work? If we use the law to evaluate evidence, will we be led to the truth? Suppose A is actually false and B is true. Can we obtain observations that, according to the law, are evidence for A over B? Certainly. Does this mean that the law is invalid? Certainly not. Evidence, properly interpreted, can be misleading. This must be the case, for otherwise we would be able to determine the truth (with perfect certainty) from any scrap of evidence that it not utterly ambiguous. It is too much to hope that evidence cannot be misleading. However, we might reasonably expect that strong evidence cannot be misleading very often. We might also expect that, as evidence accumulates, it will tend to favor a true hypothesis over a false one more and more strongly. These expectations are met by the concept of evidence embodied in the law of likelihood, as explained below.

Suppose A implies that X has probability distribution $p_A(\cdot)$, while B implies $p_B(\cdot)$. If B is true then when we observe X it is unlikely that we will find strong evidence favoring the false hypothesis A. Specifically, for any given constant $k > 0$,

$$\Pr(p_A(X)/p_B(X) \geq k) \leq 1/k. \qquad (1.2)$$

This is because, if S is the set of values of x that produce a likelihood ratio (in favor of A versus B) of at least k, then when B is correct

$$\Pr(S) = \sum_S p_B(x) \leq \sum_S p_A(x)/k \leq 1/k.$$

The first inequality is obtained because, for every x in S, $p_B(x) \leq p_A(x)/k$, and the second because the sum $\sum_S p_A(x)$ is the probability of S when A is correct, which cannot exceed one.

A similar argument can be used to prove a much stronger result: if an unscrupulous researcher sets out deliberately to find evidence supporting his favorite but erroneous hypothesis (A) over his rival's (B), which happens to be correct, by a factor of at least k, then the chances are good that he will be eternally frustrated. Specifically, suppose that he observes a sequence $X_1, X_2, \ldots$ of independent random variables, identically distributed according to $p_B(\cdot)$. He checks after each observation to see whether his accumulated data are 'satisfactory' (likelihood ratio favors A by at least k), stopping and publishing his results only when this occurs. After n observations the likelihood ratio is $\prod_1^n p_A(x_i)/p_B(x_i)$. It is a remarkable fact that the probability that he will be successful is no greater than $1/k$, and this remains true even if the number of observations he can make is limitless. That is, when B is true,

$$\Pr\left(\prod_1^n p_A(X_i)/p_B(X_i) \geq k \text{ for some } n = 1, 2, \ldots\right) \leq 1/k \qquad (1.3)$$

(Robbins, 1970).

In a more positive vein, the law of likelihood, together with the law of large numbers, implies that the accumulating evidence represented by observations on a sequence $X_1, X_2, \ldots$ of independent random variables will eventually strongly favor the truth. Specifically, if the X_i are identically distributed according to p_B, and if p_A identifies any other probability distribution, then the likelihood ratio $\prod_1^n p_A(X_i)/p_B(X_i)$ converges to zero with probability one (Exercise 1.3). This means that we can specify any large number k with perfect certainty that our evidence will favor B over A by at least k if only we take enough observations. The truth will appear. It also implies that along with k we can specify any small number $\varepsilon > 0$, then find a sample size n that will ensure that the probability of finding strong evidence (a likelihood ratio of at least k) supporting B over A is at least $1 - \varepsilon$.

1.5 Relativity of evidence

The law of likelihood applies to pairs of hypotheses, telling when a given set of observations is evidence for one versus the other: hypothesis A is better supported than B if A implies a greater probability for the observations than B does. This law represents a concept of evidence that is essentially relative, one that does not apply to a single hypothesis, taken alone. Thus is explains how observations should be interpreted as evidence for A *vis-à-vis* B, but it makes no mention of how those observations should be interpreted as evidence in relation to A alone.

When there are probabilities, $\Pr(A)$ and $\Pr(B)$, for the hypotheses, the law of likelihood implies that an observation $X = x$ that supports A over B increases the relative probability of A, as expression (1.3) shows. This observation does not necessarily increase the absolute probability of A, however. In fact, an observation that supports A over B can reduce the probabilities of both hypotheses. For example, suppose that there is another hypothesis C and that a priori $\Pr(A) = \Pr(B) = \Pr(C) = \frac{1}{3}$. If $p_A(x) = \frac{1}{6}$, $p_B(x) = \frac{1}{12}$, and $p_C(x) = \frac{1}{3}$, then the effect of the observation $X = x$ is to reduce the probability of A and of B while doubling the probability of A relative to that of B. That is, $\Pr(A|X = x) < \Pr(A)$ and $\Pr(B|X = x) < \Pr(B)$, yet

$$\frac{\Pr(A|X = x)}{\Pr(B|X = x)} = 2\frac{\Pr(A)}{\Pr(B)}.$$

The observation is not evidence supporting A taken alone – it is evidence supporting A over B. Likewise, observations can support A over B while increasing both probabilities, and such observations are evidence against B *vis-à-vis* A, but not evidence against B by itself.

Can a valid rule be found that will guide the interpretation of statistical data as evidence relating to a single hypothesis, without reference to an alternative? We will examine two candidates. The first we call the law of improbability. It states that $X = x$ is evidence against A if $p_A(x)$ is small, that is, if A implies that the observation is improbable. The second, which we call the law of changing probability, states that $X = x$ is evidence for or against A according to whether the effect of the observation is to increase or reduce the probability that A is true.

We will argue that neither of these rules represents a satisfactory concept of evidence for scientific discourse, the first because it is wrong, and the second because it is subjective. The first rule, the law of improbability, has had a powerful influence on statistical thinking. It is often cited as the justification for the popular statistical procedures called tests of significance. It will be considered in detail in Chapter 3, where we examine the rationale for tests of significance and argue that the law of improbability is wrong. The second rule, although stated in terms of a single hypothesis, and not referring to any explicit alternative, actually entails both alternative hypotheses and conditions on how prior probability is distributed among the hypotheses. Although it has had little direct impact on statistical thinking, this rule has received much attention from philosophers (Carnap, 1950; Good, 1962; Salmon, 1983).

The law of changing probability says that:

(i) the observation $X = x$ is evidence supporting A if its effect is to increase the probability of A; that is, $X = x$ supports A if $\Pr(A|X = x) > \Pr(A)$; and
(ii) the ratio $\Pr(A|X = x)/\Pr(A)$ measures the strength of the evidence.

In our diagnostic test example, the law of changing probability says that a positive test is evidence supporting the hypotheses that D is present by the factor $\Pr(D|+)/\Pr(D) = \Pr(+|D)/\Pr(+) = 0.95/\Pr(+)$. To calculate this quantity, we must know $\Pr(D)$, in which case $\Pr(+) = 0.95\Pr(D) + 0.02(1 - \Pr(D))$ and

$$\Pr(D|+)/\Pr(D) = r/[r\Pr(\mathrm{D}) + (1 - \Pr(\mathrm{D}))] \qquad (1.4)$$

where $r = 0.95/0.02$ is the likelihood ratio, $\Pr(+|D)/\Pr(+|\text{not-}D)$. Expression (1.4) is a strictly increasing function of r which equals one when $r = 1$. Thus in this case, according to the law of changing probability, the observation (+) supports D over not-D if and only if the likelihood ratio is greater than one. And the greater the likelihood ratio, the stronger the evidence. This conclusion differs from that implied by the law of likelihood only in that the measure of the evidence's strength depends on $\Pr(D)$ as well as on the ratio $\Pr(+|D)/\Pr(+|\text{not-}D)$; a given likelihood ratio in favor of D is interpreted as stronger evidence when $\Pr(D)$ is small than when this probability is large. If you and I hold different initial values for the probability of D, then we will agree that a positive test is evidence for D, but we will disagree about the strength of that evidence.

However, if the possibilities are richer, then the law of changing probability implies that we need not agree even as to the direction of the support, in favor of disease or against it. This is because, although the law of changing probability appears to measure the absolute evidence for or against hypothesis A, not the evidence for A relative to another hypothesis, this measure is in fact strongly dependent not only on what alternatives to A are considered, but also on the way a priori probabilities are distributed over the alternatives.

Suppose we want to evaluate an observation $X = x$ as evidence relating to hypothesis A, for which we know both the a priori probability of A, $\Pr(A)$, and the probability that $X = x$ if A is true, $p_A(x)$. To apply the law of changing probability we must evaluate $\Pr(A|X = x)/\Pr(A) = p_A(x)/\Pr(X = x)$, and the denominator, $\Pr(X = x)$, depends directly on alternatives to A and their a priori probabilities, as well as on the probabilities that $X = x$ under the various alternatives. For example, if there are only three possible hypotheses, A, B, and C, which have respective a priori probabilities $\Pr(A)$, $\Pr(B)$, and $\Pr(C)$, and which imply respective probabilities $p_A(x)$, $p_B(x)$, and $p_C(x)$ for the event $X = x$, then $\Pr(X = x) = p_A(x)\Pr(A) + p_B(x)\Pr(B) + p_C(x)\Pr(C)$. According to the law of changing probability the evidence for A in $X = x$ is

$$\frac{\Pr(A|X = x)}{\Pr(A)} = \frac{p_A(x)}{\Pr(X = x)}$$

$$= \frac{p_A(x)}{p_A(x)\Pr(A) + p_B(x)\Pr(B) + p_C(x)\Pr(C)}.$$

Not only does this quantity depend on the specific alternatives, B and C, that are considered (and the probabilities of $X = x$ under those alternatives), it also depends on how the a priori probability of not-A is divided between B and C. If $\Pr(A)$ is small and if $p_B(x) < p_A(x) < p_C(x)$, then the effect of the observation $X = x$ will be to increase the probability of A if $\Pr(B)$ is large, but to decrease it if $\Pr(C)$ is large. Whereas the law of likelihood measures the support for one hypothesis A relative to a specific alternative B, without regard either to the prior probabilities of the two hypotheses or to what other hypotheses might also be considered, the law of changing probability measures support for A relative to a specific prior probability distribution over A and its alternatives – the

alternatives *and* their a priori probabilities are essential to the law of changing probability, although the formula $\Pr(A|X = x)/\Pr(A)$ conceals this dependence.

The law of changing probability is of limited usefulness in scientific discourse because of its dependence on the prior probability distribution, which is generally unknown and/or personal. Although you and I agree (on the basis of the law of likelihood) that given evidence supports A over B, and C over both A and B, we might disagree about whether it is evidence supporting A (on the basis of the law of changing probability) purely on the basis of our different judgements of the a priori probabilities of A, B, and C.

1.6 Strength of evidence

How strong is the evidence when the likelihood ratio is 2? ... Or 20? Many scientists (and journal editors) are comfortable interpreting a statistical significance level of 0.05 to mean that the observations are 'pretty strong evidence' against the null hypothesis, and a level of 0.01 to mean 'very strong evidence'. Are there reference values of likelihood ratios where corresponding interpretations are appropriate? (Later, in Chapter 3, we will show that these interpretations of significance levels are not appropriate.)

There are two easy ways to develop a quantitative understanding of likelihood ratios. One is to consider some uncomplicated examples where intuition is strong, and examine the likelihood ratios for various imagined observations. The other is to characterize likelihood ratios in terms of their impact on prior probabilities.

1.6.1 A canonical experiment

Suppose we have two identical urns, one containing only white balls, and the other containing equal numbers of white and black balls. One urn is chosen and we draw a succession of balls from it, after each draw returning the ball to the urn and thoroughly mixing the contents. We have two hypotheses about the contents of the chosen urn, 'all white' and 'half white', and the observations are evidence.

Suppose you draw a ball and it is white. Suppose you draw again, and again it is white. If the same thing happens on the third draw, many would characterize these three observations as 'pretty strong' evidence for the 'all white' urn versus the 'half white' one. The likelihood ratio is $2^3 = 8$.

Table 1.2 *Number of successive white balls (b) corresponding to values of a likelihood ratio (LR)*

LR	10	20	50	100	1000
b	3.3	4.3	5.6	6.6	10.0

If we observe b successive white balls, then the likelihood ratio in favor of 'all white' over 'half white' equals $1/(\frac{1}{2})^b$, or 2^b. A likelihood ratio of 2 measures the evidence obtained on a single draw when a white ball is observed. If you would consider that observing white balls on each of three draws is 'pretty strong' evidence in favor of 'all white' over 'half white', then a likelihood ratio of 8 is pretty strong evidence.

For interpreting likelihood ratios in other problems it is useful to convert them to hypothetical numbers of white balls (Table 1.2): a likelihood ratio of k corresponds to b white balls, where $k = 2^b$, or $b = \ln k/\ln 2$. Thus if you have observations giving a likelihood ratio $p_A(x)/p_B(x) = 20$, then you have evidence favoring A over B of the same strength as the evidence favoring 'all white' over 'half white' in $b = \ln 20/\ln 2 = 4.3$ consecutive white balls (stronger than four white balls, but not as strong as five). In the diagnostic test of section 1.2 a positive result, with a likelihood ratio of 47.5, is evidence supporting D over not-D of the same strength as that favoring the 'all white' urn when $b = 5.6$ consecutive white balls are drawn.

1.6.2 Effects of likelihood ratios

Some find the preceding statements dubious. To them it is not clear that a likelihood ratio of 4, say, represents the same strength of evidence in all contexts. These doubts come from failure to distinguish between the strength of the evidence, which is constant, and its implications, which vary according to the context of each application (prior beliefs, available actions, etc.).

The key point is that observations with a likelihood ratio of 4 are evidence strong enough to quadruple a prior probability ratio. The values of the prior probabilities do not matter, nor does their ratio. The effect is always the same: a likelihood ratio of 4 produces a fourfold increase in the probability ratio. There are no circumstances where the effect is different – say, where a likelihood ratio of 4 produces a threefold or a fivefold increase. Bayes's formula guarantees this:

$$\frac{\Pr(A|X=x)}{\Pr(B|X=x)} = \frac{p_A(x)}{p_B(x)}\,\frac{\Pr(A)}{\Pr(B)}.$$

Whether the prior probabilities are known or not makes no difference; their ratio, whatever it might be, would be increased k-fold by observations with a likelihood ratio of $p_A(x)/p_B(x) = k$.

Some people are willing to state probabilities for all sorts of hypotheses, while others find it meaningful to speak of 'the probability that H is true' only for some very special hypotheses, such as those in the urn example, 'all white' and 'half white' when I choose the urn by a coin toss. The numerical value of the likelihood ratio, which is given a precise interpretation in this last situation (via Bayes's theorem), retains that meaning more generally: a likelihood ratio of k corresponds to evidence strong enough to cause a k-fold increase in a prior probability ratio, regardless of whether a prior ratio is actually available in a specific problem or not. The situation is analogous to that in physics where a unit of thermal energy, the BTU, is given a concrete meaning in terms of water – one BTU is that amount of energy required to raise the temperature of one pound of water at 39.2°F by 1°F. But it is meaningful to measure thermal energy in BTUs in rating air conditioners and in other situations where there is no water at 39.2°F to be heated. Likewise the likelihood ratio, given a concrete meaning in terms of prior probabilities, retains that meaning in their absence.

1.7 Counterexamples

We have seen that the law of likelihood is intuitively attractive; that in special situations where we know how to interpret evidence precisely (via its effect on the probabilities of hypotheses), the law is consistent with what we know to be correct; and that it works. We must test it further by examining its implications, but we will first inspect two examples which have convinced some that the law is false. Another purported counterexample is considered in section 1.10.

1.7.1 A trick deck?

I shuffle an ordinary-looking deck of playing cards and turn over the top card. It is the ace of diamonds. According to the law of likelihood, the hypothesis that the deck consists of 52 aces of diamonds

(H_1) is better supported than the hypothesis that the deck is normal (H_N) by the factor $\Pr(A\blacklozenge|H_1)/\Pr(A\blacklozenge|H_N) = 52$. (In comparison with the urn example, this is stronger than the evidence favoring 'all white' over 'half white' when five consecutive draws produce white balls.)

Some find this disturbing. Although the evidence is supposed to be strong, they would not be convinced that there are 52 aces of diamonds instead of a normal deck. Furthermore, it seems unfair; no matter what card is drawn, the law implies that the corresponding trick-deck hypothesis (52 cards just like the one drawn) is better supported than the normal-deck hypothesis. Thus even if the deck is normal we will always claim to have found strong evidence that it is not.

The first point rests on confusing evidence and belief (questions 3 and 1 in section 1.3). If drawing an ace of diamonds does not convince you that the deck has 52 aces of diamonds (H_1), this does not mean that the observation is not strong evidence in favor of H_1 versus H_N. It means simply that the evidence is not strong enough to overcome the prior improbability of H_1 relative to H_N. Edwards (1970) highlighted the role of prior opinion in our reaction to this example by considering how an individual with a somewhat different background might react:

> A Martian faced with this problem would find the first hypothesis [H_1] most appealing; are not all the cards identical in size and shape, with identical patterns on the side exposed to view? How natural, then, that they should all have the same design on the other side.

The interplanetary perspective is not necessary; we can change our own prior beliefs and see how this changes the example. Suppose I show you two decks, one normal and one actually composed of 52 aces of diamonds. I choose a deck by a coin toss, shuffle the chosen deck, and draw one card. It is an ace of diamonds. Now the conclusion that the deck is not the normal one looks quite reasonable. The evidence represented by the ace of diamonds is the same as before; it is the prior probabilities that have changed. Now the beliefs after seeing the evidence are dominated by that evidence, whereas before they were dominated by the prior beliefs.

The second objection to the law, that H_N is treated unfairly, rests on a misinterpretation: 'evidence supporting H_1 over H_N' is not 'evidence against H_N'. Consider the 51 additional different trick-deck hypotheses, $H_2, \ldots, H_{52}$, one stating that all 52 cards are fours of clubs, etc. Observing the ace of diamonds is evidence

supporting H_1 over H_N. It is also evidence supporting H_N over H_2, H_3, etc., decisively. It is not evidence for or against H_N alone.

As is often true, a Bayesian calculation can help to clarify the point. Suppose that there is some prior probability π that the deck is normal, and that if it is not normal, then it must be one of the 52 trick decks, all of which are equally probable. Thus $\Pr(H_N) = \pi$ and $\Pr(H_j) = (1-\pi)/52$, for $j = 1, 2, \ldots, 52$. How are these probabilities changed by the observation of an ace of diamonds? Bayes's theorem reveals that

$$\Pr(H_N|A\blacklozenge) = \pi,$$
$$\Pr(H_1|A\blacklozenge) = 1 - \pi,$$
$$\Pr(H_j|A\blacklozenge) = 0, \qquad j = 2, 3, \ldots, 52.$$

The probability of H_N is unchanged by the observation; the probability of H_1 is increased by a factor of 52, while the probabilities of all the other trick-deck hypotheses are driven to zero. The entire probability, $1 - \pi$, that was distributed evenly over the 52 trick-deck hypotheses is now concentrated on H_1. The probability ratio of H_1 to H_N has increased sharply, from $(1-\pi)/52\pi$ to $(1-\pi)/\pi$. But if $\pi > \frac{1}{2}$, this ratio is still less than one and the normal deck remains the more probable.

This example shows the importance, as well as the difficulty, of maintaining the critical distinction between *evidence* and *confidence* (degree of belief). The next example makes a similar point.

1.7.2 Greater confidence without stronger evidence?

Suppose that two distributions, labelled θ_1 and θ_2, both assign the same probability to a specific outcome x – say, $f(x;\theta_1) = f(x;\theta_2) = 1/20$. The hypotheses H_1: $\theta = \theta_1$ and H_2: $\theta = \theta_2$ both imply that the event $X = x$ has probability 1/20, so that, according to the law of likelihood, the occurrence of this event is not evidence supporting either hypothesis over the other.

Now consider the composite hypothesis: H_C: $\theta = \theta_1$ or θ_2. Because this hypothesis also implies that the event $X = x$ has probability 1/20, the same as the probability under H_1, the law asserts that this evidence does not support H_C over H_1 – these two hypotheses are equally well supported. But H_C must be true if H_1 is; therefore H_C is more likely to be true, more plausible, more believable, more tenable than H_1. Does this not imply that the evidence really does support H_C over H_1, contrary to the law?

There are two pieces of evidence here. One is statistical, the observation, $X = x$. The other is logical, the relationship between the two hypotheses. This second bit of evidence implies that H_C is the more credible hypothesis, independently of the first. It does not imply that the statistical evidence supports H_C over H_1. On the other hand, the law of likelihood addresses only the statistical evidence, not that in the logical relationship between the hypotheses. There is no inconsistency in acknowledging both that H_C is more credible than H_1 (because of their logical relationship, and independently of the observation) and that the observation $X = x$ is evidence favoring neither.

1.8 Testing simple hypotheses

According to the law of likelihood, the strength of statistical evidence for one hypothesis *vis-à-vis* another is measured by the likelihood ratio. This ratio also plays a central role in the Neyman–Pearson theory of hypothesis testing, but that theory addresses a different problem than the law of likelihood does. Neyman–Pearson theory, which will be discussed in Chapter 2, is aimed at using the observations to choose between two hypotheses, H_1 and H_2, not at representing or interpreting the observations as evidence. The choice is made as follows. Before X is observed a set R of possible values of X is selected. This set is called the **critical region**. Then if $X = x$ is observed and x is in R, H_2 is chosen; if x is not in R, H_1 is chosen.

Neyman and Pearson (1933) pointed out that two types of error can be made: if H_1 is true then an observation in R will lead to erroneous choice of H_2, a Type I error; a Type II error occurs when H_2 is true but the observation is not in R, so that H_1 is chosen. The probability of a Type I error is called the **size** of the critical region and is denoted by α.

For the case of simple hypotheses, H_1 and H_2, Neyman and Pearson sought, among all critical regions whose size does not exceed a specified value, such as $\alpha \le 0.05$, the one that has the smallest Type II error probability. They discovered that this best critical region is determined by the likelihood ratio. It is $R = \{x: f_2(x)/f_1(x) \ge k\}$. That is, the best test procedure is 'Choose H_2 if the likelihood ratio is at least k' where k is chosen to give the desired size, α.

For example, suppose the hypotheses specify different values for the success probability in 30 *Bernoulli*(θ) trials – say, $H_1: \theta = \frac{1}{4}$ and $H_2: \theta = \frac{3}{4}$. When the number of successes observed is x, the likelihood ratio in favor of H_2 over H_1 is $f_2(x)/f_1(x) = 3^{2x-30}$. The best critical region with size $\alpha = 0.05$ contains all values of x for which the likelihood ratio is at least $k = 3^{24-30} = \frac{1}{729}$, that is, $x \ge 12$. (Under H_1 the probability of 12 or more successes in 20 trials is only 0.05.)

It is reassuring to find that the best test calls for choosing H_2 when evidence favors H_2 over H_1 by a sufficiently large factor (k). But, as this example shows, the critical factor k can be less than one, and in that case the test sometimes calls for choosing H_2 when the evidence actually favors H_1. For instance, when $x = 12$ is observed, the test calls for choosing H_2, although the observation is strong evidence supporting H_1 over H_2 ($f_1(12)/f_2(12) = 729$, evidence stronger than when nine consecutive white balls are drawn in the urn example of section 1.6).

Similarly, the observations $x = 13$ and 14 are fairly strong evidence in favor of H_1. And the observation $x = 15$, which represents a success rate of $\frac{1}{2}$, equally far from the two hypothesized values, $\theta = \frac{1}{4}$ and $\theta = \frac{3}{4}$, is not evidence supporting H_2 over H_1, but utterly neutral evidence ($f_2(15)/f_1(15) = 1$).

Although likelihood theory and Neyman–Pearson testing theory have much in common, it is clear that there are fundamental differences. While likelihood theory addresses the last of the physician's three questions in Chapter 1 (What does this observation say about H_1 versus H_2?), Neyman–Pearson theory is concerned with the second question (What should I do?). It is interesting to note that the link between the two theories would be even stronger if, instead of minimizing the probability of a Type II error for a fixed value of α, Neyman and Pearson had sought to minimize the sum of the two error probabilities; in that case they would have found that the best critical region consists of those observations whose likelihood ratio is greater than one. That is, they would have found that the best rule is to choose the hypothesis that is better supported by the observations (Exercise 1.6; see Cornfield, 1966). We will take a closer look at the Neyman–Pearson statistical theory in Chapter 2.

1.9 Composite hypotheses

The law of likelihood explains how an observation on a random variable should be interpreted as evidence in relation to two simple statistical hypotheses. It also applies to some composite hypotheses, such as H_C in section 1.7. But it does not apply to composite hypotheses generally. A simple example shows why this is so.

Suppose three probability distributions, labelled θ_1, θ_2, and θ_3, are under consideration for a random variable X. In particular, we want to evaluate an observation $X = x$ as evidence for the simple hypothesis H_2: $\theta = \theta_2$ *vis-à-vis* the composite H_C: $\theta = \theta_1$ or θ_3. Suppose $f(x;\theta_1) > f(x;\theta_2) > f(x;\theta_3)$, that is, the observation is evidence supporting H_1: $\theta = \theta_1$ over H_2, but it also supports H_2 over H_3: $\theta = \theta_3$.

For example, X might be the number of white balls in five draws (with replacement) from an urn whose proportion of white balls is either one-fourth (θ_1), one-half (θ_2) or three-fourths (θ_3). If none of the five draws produces a white ball ($X = 0$), this is evidence supporting H_1 over H_2 by a factor of $(\frac{3}{4})^5/(\frac{1}{2})^5 = \frac{243}{32}$, or about 7.6. But it also supports H_2 over H_3, by $(\frac{1}{2})^5/(\frac{1}{4})^5 = 32$. How about H_C (proportion is either $\frac{1}{4}$ or $\frac{3}{4}$) versus H_2? Because H_C does not imply a definite probability for the observation, the law of likelihood is silent.

Perhaps the law as stated above is unnecessarily restricted, and an acceptable extension might be found which would imply that $X = 0$ is evidence for H_C versus H_2. One argument that might be advanced in support of this speculation is as follows: H_C must be true if H_1 is true, and H_1 is supported over H_2. Thus it might seem reasonable to say that H_C is also supported over H_2. But this argument rests on the same fallacy as the one at the end of section 1.7 – confusing what the logical structure implies with what the statistical data tell us.

Examination of the evidence's effect on the relative probabilities of H_C and H_2, quantified by Bayes's formula, confirms that the suggested extension of the law of likelihood goes too far: if the three values θ_1, θ_2, and θ_3 have respective prior probabilities p_1, p_2, and p_3, then

$$\frac{\Pr(H_C|X=0)}{\Pr(H_2|X=0)} = \frac{(243/32)p_1 + (1/32)p_3}{p_2}.$$

This ratio is larger or smaller than the a priori probability ratio, $(p_1 + p_3)/p_2$, according to whether p_1/p_3 is larger or smaller than 31/211. Observation of no white balls in five draws causes an increase in the probability of H_C, compared to that of H_2, if the ratio of prior probabilities of the components of H_C, p_1/p_3, is large enough, and otherwise causes a decrease.

More generally, let r_{12} be the likelihood ratio of H_1 to H_2, $f(x;\theta_1)/f(x;\theta_2)$, and let r_{32} be the ratio of H_3 to H_2. Then

$$\frac{\Pr(H_C|X=x)}{\Pr(H_2|X=x)} = \frac{r_{12}p_1 + r_{32}p_3}{p_2},$$

which equals

$$\frac{[r_{12}w + r_{32}(1-w)](p_1+p_3)}{p_2},$$

where $w = p_1/(p_1+p_3)$. Thus the term in square brackets is the factor by which the prior probability ratio $\Pr(H_C)/\Pr(H_2)$ is altered by the observation $X = x$. This factor is a weighted average of the two likelihood ratios, r_{12} and r_{32}, the weights being the proportions of the total probability of H_C, $p_1 + p_3$, that are assigned to the respective components, H_1 and H_3. Since this factor is at least as large as the smaller of the two likelihood ratios, we can properly characterize the observation as evidence favoring H_C over H_2 when both r_{12} and r_{32} exceed unity, that is, when the evidence supports each of the components of H_C, H_1 and H_3, over H_2. But when $r_{12} > 1 > r_{32}$ we can interpret the evidence for H_C versus H_2 only in relation to the price probability ratio p_1/p_3.

What, then, can we say about statistical evidence when many more than two simple probability distributions are of genuine interest? Does the law of likelihood (without prior probabilities) provide a means of representing and measuring the evidence which is appropriate for scientific interpretation and communication? To show that it does, we consider an example that is examined in greater detail in section 6.2. In that example medical researchers are interested in the success probability, θ, associated with a new treatment. They are particularly interested in how θ relates to the old treatment's success probability, believed to be about 0.2. They have reason to hope that θ is considerably greater, perhaps 0.8 or even greater. To obtain evidence about θ, they carry out a study in which the new treatment is given to 17 subjects, and find that it is successful in nine.

A standard statistical analysis of their observations would use a *Bernoulli*(θ) statistical model and test the composite hypotheses H_1: $\theta \leq 0.2$ versus H_2: $\theta > 0.2$. That analysis would show that H_1 can be rejected in favor of H_2 at any significance level greater than 0.003, a result that is conventionally taken to mean that the observations are very strong evidence supporting H_2 over H_1.

But because H_1 contains some simple hypotheses that are better supported than some hypotheses in H_2 (e.g. $\theta = 0.2$ is better supported than $\theta = 0.9$ by a likelihood ratio of $LR = (0.2/0.9)^9(0.8/0.1)^8 = 22.2$), the law of likelihood does not allow the characterization of these observations as strong evidence for H_2 over H_1.

What does it allow us to say? One statement that we can make is that the observations are only weak evidence in favor of $\theta = 0.8$ versus $\theta = 0.2$ ($LR = 4$). We can also say that they are rather strong evidence supporting $\theta = 0.5$ over any of the values under $H_1: \theta \leq 0.2$ ($LR > 89$), and at least moderately strong evidence for $\theta = 0.5$ over any value $\theta \geq 0.8$ ($LR > 22$). These ratios change very little if we replace $\theta = 0.5$ by slightly different values. Thus we can say that the observation of nine successes in 17 trials is rather strong evidence supporting success rates of about 0.5 over the rate 0.2 that is associated with the old treatment, and at least moderately strong evidence for the intermediate rates versus the rates of 0.8 or greater that we were hoping to achieve.

The law of likelihood does not allow us to characterize the evidence in terms of the hypotheses $H_1: \theta \leq 0.2$ and $H_1: \theta > 0.2$. It forces us to be more specific, to note and report which values greater than 0.2 are better supported than values of 0.2 or less, for example, and by how much. As we will see later (Figure 1.1), the law of likelihood enables us to see, understand, and communicate the evidence as it pertains not just to two pre-selected hypotheses, but to the totality of possible values of θ.

1.10 Another counterexample

I have written numbers on two cards. On one card I wrote the number $\pi^{-1} = 0.318$ and on the other I wrote the value of a standard normal deviate (recorded to three decimal places). One of the cards is lost, and I am curious about the remaining one, which is in my desk drawer. Is it the card on which I deliberately wrote 0.318, or is it the one with the random number? Here we have two simple hypotheses; one, H_D, states that the number on the card, X, equals 0.318 (with probability one), and the other, H_N, states that X has a standard normal probability distribution. If I open the drawer and observe the value of X, I will have evidence concerning these two hypotheses, and if that value is 0.318, it is evidence supporting H_D over H_N by a very large factor. (Since I rounded the normal deviate, the probability of $X = 0.318$ under H_N is approximately $0.001\phi(0.318) = 1/2637$, where ϕ is the standard normal probability density function. Of course, the probability under H_D is one, so the likelihood ratio in favor of H_D, $P_D(X = 0.318)/P_N(X = 0.318)$, is approximately 2637. This is very strong evidence, having about the same strength as that supporting the hypothesis 'all white' balls versus 'half white' balls in the urn example of section 1.6 when we draw 11 consecutive white balls.

This interpretation of the evidence, guided by the law of likelihood, seems entirely reasonable. I know of no arguments to the contrary. In fact, this example has never been cited as a counterexample to the law. But it bears directly on the next one, which has (Hacking, 1972; Hill, 1973; Cox and Hinkley, 1974, p. 52; see also Birnbaum, 1969, pp. 127–8).

Supposing that X has a normal distribution, consider the evidence in the single observation $X = x$. The likelihood ratio for comparing the evidence for simple hypotheses H_1: $N(\mu_1, \sigma_1^2)$ and H_2: $N(\mu_2, \sigma_2^2)$ is

$$\frac{\sigma_2}{\sigma_1}\exp\frac{1}{2}\left[\left(\frac{x-\mu_2}{\sigma_2}\right)^2 - \left(\frac{x-\mu_1}{\sigma_1}\right)^2\right],$$

which increases without limit as (μ_1, σ_1^2) approach the values $(x, 0)$. This means that, regardless of what the true values of (μ, σ^2) are, so long as σ^2 is not zero, we will always find very strong evidence in favor of another hypothesis, namely that $\mu = x$ and $\sigma^2 = 0$. Uneasiness with this conclusion appears to grow from its misinterpretation, of which there are at least three varieties:

1. The conclusion means that no matter what is observed, it will be strong evidence against the true hypothesis (but this seems both unfair and incorrect).
2. The conclusion means that whatever value x is observed, I should be moved to attach a high degree of belief to the hypothesis $N(x, 0)$, but I am not so moved.
3. The conclusion means that the evidence is strong that σ^2 is very small (but it is intuitively clear that with no prior information about μ, one observation can provide no evidence about σ^2).

The first two misinterpretations are analogous to those already discovered in the earlier example of one card drawn from a well-shuffled deck. The error involved in the first is that it overlooks the relativity of evidence: the fact that we can find some other hypothesis that is better supported than H does not mean that the observations are evidence against H. Reaction 2 was also explained in the same example: evidence must not be confused with belief. Whether or not I am moved to attach a high degree of belief to $N(x,0)$ depends on my prior belief in that hypothesis. If I remain skeptical, it does not show that the evidence does not favor $N(x,0)$ over every other hypothesis. What it does show is that my prior skepticism was so strong that it is not overwhelmed by this evidence. As the previous example with $x = 0.318$ showed, if $N(x,0)$ had been a hypothesis that was plausible before X was observed, then the observation $X = x$ would have elicited a high degree of confidence in its truth.

The third interpretation, that $X = x$ is strong evidence that σ^2 is small, has many facets. One problem is that 'σ^2 is small', or even the more restrictive statement '$\sigma^2 = 0$', is a composite hypothesis, allowing the other parameter μ to range over the entire real line. Concerning the relative support for the composite hypotheses, $\sigma^2 = 0$, versus $\sigma^2 = \sigma_0^2$, say, the law is silent. A claim that the former is the better supported must rest on some additional principle or convention; it is not sanctioned by the law of likelihood. The problem of evaluating evidence concerning one parameter in models that also contain other parameters ('nuisance' parameters) is one with no general solution. The cause has already been described in section 1.9. We address this problem in Chapters 6 and 7, where we will see that for most problems there are quite satisfactory *ad hoc* methods for representing and interpreting evidence in the presence of nuisance parameters.

1.11 Irrelevance of the sample space

The law of likelihood says that the evidence in an observation, $X = x$, as it pertains to two probability distributions labelled θ_1 and θ_2, is represented by the likelihood ratio, $f(x;\theta_1)/f(x;\theta_2)$. In particular, the law implies that for interpreting the observation as evidence for hypothesis $H_1\colon \theta = \theta_1$ *vis-à-vis* $H_2\colon \theta = \theta_2$, only the likelihood ratio is relevant. What other values of X might have been observed, and how the two distributions in question spread their remaining probability over the unobserved values is irrelevant – all that counts is the ratio of the probabilities of the observation under the two hypotheses.

This is made clear by examples like the following (Pratt, 1961). Suppose that the hypotheses concern the probability θ of heads when a particular bent coin is tossed, and that, to generate evidence

about θ, 20 tosses are made. The result is reported in code, and you, knowing the code, will learn precisely how many tosses produced heads. I know only the code-word for '6', so that from the report I can determine only whether the outcome is '6' or 'not-6'. Thus you will observe a random variable X taking values $x = 0, 1, \ldots, 20$ with probabilities $\binom{20}{x}\theta^x(1-\theta)^{20-x}$. I will observe Y, which equals 6 with probability $\binom{20}{6}\theta^6(1-\theta)^{14}$ and some other value, c, representing the outcome 'not-6', with probability $1 - \binom{20}{6}\theta^6(1-\theta)^{14}$. Your sample space consists of the 21 points $\{0, 1, \ldots, 20\}$ while mine consists of $\{6, c\}$.

Now suppose the experiment is done and heads occur on six tosses. For any probabilities θ_1 and θ_2, my evidence concerning $H_1\colon \theta = \theta_1$ *vis-à-vis* $H_2\colon \theta = \theta_2$ is the same as yours; the likelihood ratios for $Y = 6$ and for $X = 6$ are identical: $\theta_1^6(1-\theta_1)^{14}/\theta_2^6(1-\theta_2)^{14}$. Of course, if the experiment had a different outcome, your evidence and mine would have been different; if there had been four heads you would have observed $X = 4$, and your likelihood ratio would have been $\theta_1^4(1-\theta_1)^{16}/\theta_2^4(1-\theta_2)^{16}$, while I would have observed only $Y = c$, giving a likelihood ratio of $[1 - 38\,760\theta_1^6(1-\theta_1)^{14}]/[1 - 38\,760\theta_2^6(1-\theta_2)^{14}]$, but that is irrelevant to the interpretation of the observation at hand, $X = Y = 6$. Although the scientific community might reasonably have chosen to subscribe to your newsletter in preference to mine, on the grounds that you could promise to provide a more detailed description of the observation under most circumstances, for the result that actually occurred, six heads, my report is equivalent, as evidence about θ, to yours. Any concept or technique for evaluating observations as evidence that denies this equivalence, attaching a different measure of 'significance' to your report of this result than to mine, is invalid. Whatever can be properly inferred about θ from your report can be inferred from mine and vice versa. The difference between our sample spaces is irrelevant.

We will see this example again in the following section and in section 3.4, where we use it to illustrate the problems that arise when significance tests are used to measure the strength of statistical evidence. The 'irrelevance of the sample space' is a critically important concept, for it implies a structural flaw that is not limited to significance tests, but pervades all of today's dominant statistical methodology.

1.12 The likelihood principle

Suppose two simple hypotheses for the distribution of a random variable X assign respective probabilities $f_1(x)$ and $f_2(x)$ to the outcome $X = x$, while two different hypotheses for the distribution of another random variable Y assign respective probabilities $g_1(y)$ and $g_2(y)$ to the outcome $Y = y$. If $f_1(x)/f_2(x) = g_1(y)/g_2(y)$ then the evidence in the observation $X = x$ regarding f_1 *vis-à-vis* f_2 is equivalent to that in $Y = y$ regarding g_1 *vis-à-vis* g_2. If a third distribution, f_3, is considered for X, and a third, g_3, for Y, then the two outcomes, $X = x$ and $Y = y$, are equivalent evidence concerning the respective collections of distributions, $\{f_1, f_2, f_3\}$ and $\{g_1, g_2, g_3\}$, if all of the corresponding likelihood ratios are equal: $f_1(x)/f_2(x) = g_1(y)/g_2(y), f_1(x)/f_3(x) = g_1(y)/g_3(y)$, etc. This fact is called the **likelihood principle**; it is usually stated in terms of likelihood functions, which we now define.

It is often convenient to use a parameter θ to label the individual members of a collection of probability distributions, so that each distribution is identified by a unique value of θ. The collection of distributions is $\{f(\cdot\,;\theta); \theta \in \Theta\}$, where Θ is simply the set of all values of θ. If $\theta = \theta_1$ then the probability that $X = x$ is given by $f(x;\theta_1)$. If the distributions are continuous the same notation, $f(x;\theta_1)$, represents the probability density function at the point x when the distribution is the one labelled θ_1. For a fixed value x, $f(x;\theta)$ can be viewed as a function of the variable θ and it is then called the **likelihood function**. We will use the notation $L(\theta; x)$ for the likelihood function when the value of x needs to be made explicit, and use simply $L(\theta)$ when it does not. The law of likelihood gives this function its meaning: if $L(\theta_1; x) > L(\theta_2; x)$, then the observation x is evidence supporting the hypothesis that θ is θ_1 (that is, the hypothesis that X has the distribution identified with the parameter value θ_1) over the hypothesis that θ is θ_2, and the **likelihood ratio** $L(\theta_1; x)/L(\theta_2; x) \equiv f(x;\theta_1)/f(x;\theta_2)$ measures the strength of that evidence. Because only ratios of its values are meaningful, the likelihood function is defined only up to an arbitrary multiplicative constant – $L(\theta; x) = cf(x;\theta)$.

The likelihood principle asserts that two observations that generate identical likelihood functions are equivalent as evidence; in

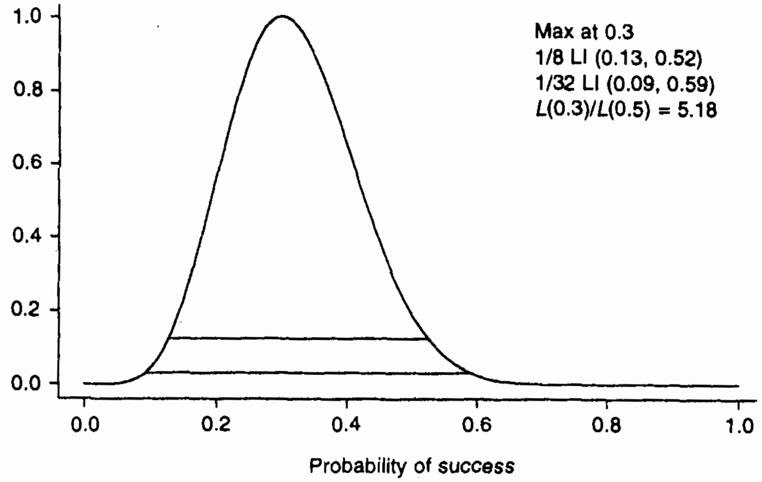

Figure 1.1 *Likelihood for probability of success: six successes observed in 20 trials.*

Birnbaum's (1962) words, 'the "evidential meaning" of experimental results is characterized fully by the likelihood function'.

The example in the previous section concerns a family that contains many more than three distributions. Your observation of the number of heads in 20 tosses of a bent coin was modelled as an observation on a random variable X with a binomial probability distribution, $Bin(20, \theta)$. The probability of six successes is $\Pr(X = 6) = \binom{20}{6}\theta^6(1-\theta)^{14}$, so the likelihood function, $L(\theta)$, is proportional to $\theta^6(1-\theta)^{14}$, for $0 \le \theta \le 1$. This function appears in Figure 1.1, which shows that its maximum is at $\theta = 6/20 = 0.3$ (the 'best-supported' hypothesis) and that $\theta = 0.3$ is better supported than $\theta = 0.5$ by a modest factor of $L(0.3)/L(0.5) = (0.3)^6(0.7)^{14}/(0.5)^{20} = 5.2$. This is slightly stronger than the evidence in favor of the 'all white' urn (section 1.6) when two white balls are drawn.

A horizontal line is drawn in Figure 1.1 to show the values of θ where the ratio of $L(\theta)$ to the maximum, $L(0.3)$, is greater than 1/8. Another line shows where it is greater than 1/32. These lines define 'likelihood intervals' (LIs) which, along with the maximizing value, provide a useful summary of what the data say under the present model. The values 1/8 and 1/32 are used because they correspond to the likelihood ratio in the urn example of section

1.6 when, respectively, three and five consecutive white balls are drawn.

Values within the 1/8 likelihood interval are those that are 'consistent with the observations' in the strong sense that there is no alternative value that is better supported by a factor greater than 8. Thus if θ is in this interval, then there is no alternative for which these observations represent 'fairly strong evidence' in favor of that alternative *vis-à-vis* θ. For any θ outside this interval there is at least one alternative value, namely the best-supported value, 0.3, that is better supported by a factor greater than 8. The 1/32 likelihood interval has the same interpretation, but with the bench-mark value 32, representing 'quite strong' evidence, replacing 8.

If I perform the same physical experiment but can only discern Y, which indicates whether the result was '6' or not, then for me the probability of the same observation, six successes, is the same, $\Pr(Y=6)=\Pr(X=6)=\binom{20}{6}\theta^6(1-\theta)^{14}$, so that my likelihood function is the same as yours. The evidence about the probability of heads is represented by that likelihood function (Figure 1.1), and it is the same in both cases – the difference between our sample spaces is irrelevant.

Suppose now that I perform an entirely different experiment. Instead of fixing the number of tosses at 20 I resolve to keep tossing until I get six heads, then to stop. The random variable is now Z, representing the number of tosses required. If I observe $Z=20$, the probability is $\Pr(Z=20)=\binom{19}{5}\theta^6(1-\theta)^{14}$, which is different from $\Pr(X=6)=\Pr(Y=6)$. But the likelihood function is the same, proportional to $\theta^6(1-\theta)^{14}$. For every pair of values, θ_1 and θ_2, the likelihood ratio is the same when $Z=20$ as when $X=Y=6$, so the results are all equivalent as evidence about θ. It is again represented in Figure 1.1 and has precisely the same interpretation in all three cases.

This conclusion calls into question analyses that use the most popular concepts and techniques in applied statistics (unbiased estimation, confidence intervals, significance tests, etc.) when these analyses purport to represent 'what the data say about θ', i.e. to convey the meaning of the observations as evidence about θ. These conventional analyses are questionable because they all certify that there are important differences between observing $X=6$, $Y=6$, and $Z=6$, whereas these three results are in fact evidentially equivalent. For instance, the observed proportion of successes is an unbiased estimator of θ in the first case, but not in the third; a 95% confidence interval for θ based on observation $X=6$ will differ from those based on $Y=6$ or $Z=20$; and for testing hypotheses about θ, e.g. H_1: $\theta\geq 0.5$ versus H_2: $\theta<0.5$, the three observations will give different p-values. We will consider this example again in Chapter 3.

This is not to say that there are not important differences between the three experiments. The first two were sure to be finished after 20 tosses, while the third could have dragged on indefinitely. The first might have produced 20 consecutive heads, giving a likelihood ratio favoring $\theta_1=3/4$ over $\theta_2=1/4$ by a factor of more than 3 billion! The second and third experiments cannot possibly generate such strong evidence in favor of θ_1 over θ_2. But the third, by producing a very large value for Z, could have provided much stronger evidence in favor of θ_2 over θ_1 than any possible outcome of the first. The experiments are certainly not equivalent; yet if the first one produces $X=6$, this outcome is equivalent, as evidence about θ, to the outcome $Y=6$ of the second and to the outcome $Z=20$ of the third.

The foregoing conclusion applies even when the parameter θ has a totally different meaning in the different experiments. If you make 20 tosses of a bent coin and observe $X=6$ heads and I count the number of cars before the sixth Ford has passed my house and observe $Z=20$ then your evidence about the probability of heads and mine about the probability of a Ford are equivalent. Of course, this equivalence only applies to the specific families of probability distribution being considered:

$$\Pr(X=x)=\binom{20}{x}\theta^x(1-\theta)^{20-x},\quad x=0,1,\ldots,10,$$

for X and

$$\Pr(Z=z)=\binom{z-1}{5}\theta^6(1-\theta)^{z-6},\quad z=6,7,\ldots,$$

for Z. The evidence supports $\theta=1/4$ over $\theta=1/10$ by the factor 19.02, regardless of whether it is your evidence and θ is the probability of heads or it is mine and θ is the probability of Fords. The evidential equivalence of your observation and mine *vis-à-vis* our respective families of distributions applies only to comparisons made within those specific families; it clearly does not assume or imply that my family is as adequate as a model for the frequency of Fords as yours is for the frequency of heads.

This concept – that results of two different experiments have the same evidential meaning if they generate the same likelihood function – has been the focus of much controversy among statisticians. Birnbaum (1962) gave a formal statement of the concept, which he called the likelihood principle. Other authors, notably Fisher (1922) and Barnard (1949), had previously promoted the concept, but most statisticians were not convinced of its validity. Birnbaum increased the pressure on the doubters by showing that the likelihood principle can be deduced from two other principles that most of them did find compelling, the sufficiency and conditionality principles. Since the publication of Birnbaum's result in 1962 statistics has struggled to understand it and to resolve the dilemma that it created (Birnbaum, 1962; 1972; 1977; Durbin, 1970; Savage, 1970; Kalbfleisch, 1975; Berger and Wolpert, 1988; Berger, 1990; Joshi, 1990; Bjornstad, 1996).

1.13 Evidence and uncertainty

We have suggested that the concept of statistical evidence is properly expressed in the law of likelihood and that the likelihood function is the appropriate mathematical representation of statistical evidence. Many likelihood functions, like the one in Figure 1.1, for example, look like probability density functions. However, there are critical differences between the two kinds of function, both in terms of what they mean and in terms of what mathematical operations make sense.

Probabilities measure uncertainty and likelihood ratios measure evidence. A probability density function represents the uncertainty about the value of a random variable; it describes how the uncertainty is distributed over the possible values of the variable (the sample space). That uncertainty disappears when the observation is made – then the value of the random variable is known, and that value is evidence about the probability distribution. The likelihood function represents this evidence; it describes the support ratio for any pair of distributions in the probability model.

Sometimes one variable appears in both aspects of a problem. It is itself a potentially observable random variable, and it is also a parameter that identifies the probability distribution of a second random variable. If (X, Y) are random variables with a given joint probability distribution, then after $X = x$ is observed, $f_{Y|X}(y|x)$ represents the uncertainty about the value of Y. (Note that if we denote the second random variable by Θ instead of Y, then we have the Bayesian statistical model and the Bayesian 'solution' to the problem of statistical inference, $f_{\Theta|X}(\theta|x)$. Bayesian statistics will be discussed in Chapter 8.) But the unobserved value y plays the role of a parameter in $f_{X|Y}(x|y)$, so that the observation $X = x$ is statistical evidence about y, generating a likelihood function $L(y) \propto f_{X|Y}(x|y)$ which represents that evidence. Comparing these two functions in a familiar example helps to clarify their differences.

Consider the case when x and y are realizations of random variables, X and Y, having a bivariate normal probability distribution with expected values μ_x and μ_y, variances σ_x^2 and σ_y^2, and covariance σ_{xy}. Suppose the values of all five parameters are known. If x and y have not yet been observed then the uncertainty about the value y is expressed in the marginal probability distribution of Y, which is $N(\mu_y, \sigma_y^2)$. The observation $X = x$ represents evidence about y. It changes the uncertainty, which, after $X = x$ is observed, is represented by the conditional probability distribution of Y, $N(\mu_y + \sigma_{xy}(x - \mu_x)/\sigma_x^2, \sigma_y^2(1 - \rho^2))$, where ρ denotes the correlation coefficient, $\sigma_{xy}/\sigma_x\sigma_y$. This probability density function, $f_{Y|X}(y|x)$, represents the uncertainty about what value, y, of Y will be observed, now that it is known that $X = x$.

On the other hand, the variable y indexes a family of possible probability distributions for X. These are the conditional distributions of X, given $Y = y$, which are $N(\mu_x + \sigma_{xy}(y - \mu_y)/\sigma_y^2, \sigma_x^2(1 - \rho^2))$. Here y has the role of a parameter – each value of y determines a different probability distribution for X, $f_{X|Y}(x|y)$. Thus the observation $X = x$ generates a likelihood function for y,

$$L(y) \propto \exp\{-\tfrac{1}{2}[x - \mu_x - \sigma_{xy}(y - \mu_y)/\sigma_y^2]^2/\sigma_x^2(1 - \rho^2)\}. \quad (1.5)$$

The only variable in this expression is y – everything else, x, μ_x, etc., is fixed at its known value. The ratio of values of this function at any two points y_1 and y_2, $L(y_1)/L(y_2)$, measures the relative support for these two values of the unknown variable y.

If X and Y are independent, so that $\sigma_{xy} = 0$, then the likelihood function (1.5) for y is a constant, indicating that $X = x$ represents no evidence at all about y. Every likelihood ratio $L(y_1)/L(y_2)$ equals one – when $\sigma_{xy} = 0$ no possible value of y is better supported than any other by the observation $X = x$, regardless of the value of x.

When X and Y are not independent the likelihood function is shaped like a normal probability density function centered at the point $\mu_y + \sigma_y^2(x - \mu_x)/\sigma_{xy}$ and with variance $\sigma_y^2(1 - \rho^2)/\rho^2$. That is, the likelihood function given in expression (1.5) can be rewritten in the form:

$$L(y) \propto \exp\left\{-\frac{1}{2}\,\frac{[y-\mu_y-\sigma_y^2(x-\mu_x)/\sigma_{xy}]^2}{\sigma_y^2(1-\rho^2)/\rho^2}\right\}.$$

This function represents the *evidence* about y in the observation $X = x$. It does not represent the *uncertainty* about y, which is now given by the conditional probability density function of Y, given $X = x$:

$$f_{Y|X}(y|x) \propto \exp\left\{-\frac{1}{2}\,\frac{[y-\mu_y-\sigma_{xy}(x-\mu_x)/\sigma_x^2]^2}{\sigma_y^2(1-\rho^2)}\right\}. \tag{1.6}$$

This density function is obtained by adjusting the original $N(\mu_y, \sigma_y^2)$ density function, $f_Y(y)$, in the light of the evidence $X = x$. The adjustment is made simply by taking the product of the original density and the likelihood function $L(y)$. To use this density function we must scale it so that its integral over the entire real line equals one. When we do that, by dividing expression (1.6) by $[2\pi\sigma_y^2(1-\rho^2)]^{1/2}$, integration over any interval then gives the probability that the value of Y will fall inside that interval. This implies, for example, that the probability is 0.95 that y will be found in the predictive interval

$$\mu_y + \sigma_{xy}(x-\mu_x)/\sigma_x^2 \pm 1.96\sigma_y(1-\rho^2)^{1/2}. \tag{1.7}$$

On the other hand, the $1/k$ likelihood interval, $\{y;\, L(y)/\max L(y) \geq 1/k\}$, which is the set of y values such that no alternative is better supported by a likelihood ratio greater than k, is

$$\mu_y + \sigma_y^2(x-\mu_x)/\sigma_{xy} \pm (2\ln k)^{1/2}(\sigma_y/\rho)(1-\rho^2)^{1/2}. \tag{1.8}$$

The probability is 0.95 that the value of the random variable Y associated with the observed value x will fall in the predictive interval (1.7). No such simple probability statement can be made about the likelihood interval (1.8). But that interval can be interpreted as a confidence interval. Suppose we use the value $k = \exp\{(1.96)^2/2\} = 6.83$, so that the coefficient $(2\ln k)^{1/2}$ equals 1.96. Then for any fixed value y of the random variable Y, if we observe the value of the random variable X and construct the interval (1.8), the probability that this random interval will contain y equals 0.95 (Exercise 1.7). The purpose here is not to suggest that likelihood intervals should be interpreted as confidence intervals, but simply to clarify the distinction between the state of uncertainty about y after observing $X = x$, which is represented by the conditional probability density function, and the evidence about y in the observation $X = x$, which is represented by the likelihood function. The distinction is essentially the same as that between the physician's first and third questions in section 1.3, here rephrased as 'What is the state of uncertainty about y, now that we know that $X = x$?' and 'What does the observation $X = x$ tell us about y?'. The probability density function $f_{Y|X}(y|x)$ answers the first question, and the likelihood function $L(y)$ answers the second. When $\sigma_{XY} = 0$, $X = x$ tells us nothing about y. This is properly represented by the flat likelihood function, $L(y) = \text{constant}$; the probability density function, $f_{Y|X}(y|x) \propto \exp\{-\frac{1}{2}(y-\mu_y)^2/\sigma_y^2\}$, represents something quite different.

1.14 Summary

The question that is at the heart of statistical inference – 'When is a given set of observations evidence supporting one hypothesized probability distribution over another?' – is answered by the law of likelihood. This law effectively defines the concept of statistical evidence to be relative, that is, a concept that applies to one distribution only in comparison to another. It measures the evidence with respect to a pair of distributions by their likelihood ratio.

The law of likelihood is intuitively reasonable, consistent with the rules of probability theory, and empirically meaningful. It is, however, incompatible with today's dominant statistical theory and methodology, which do not conform to the law's general implications, the irrelevance of the sample space and the likelihood principle, and which are articulated in terms of probabilities, which measure uncertainty, rather than likelihood ratios, which measure evidence.

Exercises

1.1 The law of likelihood is stated in section 1.2 for discrete distributions. Suppose that two hypotheses, A and B, both imply that a random variable X has a continuous probability distribution, and that these distributions have continuous density functions, $p_A(x)$ and $p_B(x)$ respectively. Can the law be extended to this case? Explain.

1.2 Suppose A implies that X has probability mass (or density) function $p_A(x)$, while B implies $p_B(x)$. When A is true, observing a value of X that represents strong evidence in favor of B ($p_B(x)/p_A(x) \geq k$) is clearly undesirable. We showed in section 1.4 that the probability of this event is small ($\leq 1/k$).

(a) What can you say about the probability of the desirable event, namely, finding strong evidence in favor of A?

(b) It would be nice if, when A is true, the probability of obtaining strong evidence in favor of A is always at least as great as the probability of (misleading) strong evidence in favor of B, that is, if

$$\Pr_A(p_B(X)/p_A(X) \geq k) \leq \Pr_A(p_A(X)/p_B(X) \geq k).$$

Give a simple example to show that this inequality need not hold.

1.3 Prove the result stated at the end of section 1.4, namely, if A implies that $X_1, X_2, \ldots$ are independent and identically distributed with probability mass function $p_A(x)$, while B implies the same, but with a different mass function given by $p_B(x)$, then when B is true the likelihood ratio in favor of A converges to zero with probability one.

1.4 Suppose $X_1, \ldots, X_n$ are independent, identically distributed random variables with a $N(\theta, \sigma^2)$ probability distribution, with σ^2 known. Consider two hypotheses, $H_0\colon \theta = 0$ and $H_1\colon \theta = \theta_1$, where $\theta_1 > 0$. If a sample is observed with $\sqrt{n}\bar{x}/\sigma = 1.645$, then the p-value, or 'attained significance level', for testing H_0 versus H_1 is 0.05. This p-value is often interpreted as meaning that the observations represent fairly strong evidence against H_0. (This interpretation will be discussed later, in section 3.4.) According to the law of likelihood the strength of the evidence depends on the value, θ_1, that is specified by H_1.

(a) For what value of θ_1 is this evidence for H_1 versus H_0 strongest?

(b) For the value of θ_1 in (a), what is the likelihood ratio, f_1/f_0?

(c) If k represents the number of consecutive draws producing white balls in the urn scheme of section 1.6, to what value of k does the likelihood ratio in (b) correspond?

(d) Discuss these results.

1.5 For the same model and hypotheses as in Exercise 1.4, suppose we choose some number $k > 1$ and interpret observations as strong evidence in favor of H_1 over H_0 when f_1/f_0 exceeds k.

(a) What value of θ_1 maximizes the probability, when H_0 is true, of finding strong evidence in favor of H_1?

(b) What is the maximum probability in (a)?

(c) Compare the bound in (b) with the universal bound, $1/k$, that was derived in section 1.4.

(d) When $\theta = 0$, what is the probability that for some value of $\theta_1 > 0$ the hypothesis $H_1\colon \theta = \theta_1$ will be better supported than H_0 by a factor of at least k? That is, what is the probability of observing values $x_1, x_2, \ldots, x_n$ for which some positive θ can be found that is better supported than $\theta = 0$?

1.6 Consider testing $H_1\colon X \sim f_1$ versus $H_2\colon X \sim f_2$ on the basis of an observation on X, with the goal of minimizing the sum of the two error probabilities, $\alpha + \beta$. Show that the best test procedure is to 'choose H_2 if the observation is evidence supporting H_2 over H_1; otherwise choose H_1'. The critical region that corresponds to this rule is $R_0 = \{x; f_2(x) > f_1(x)\}$. Show also that the critical region $R_1 = \{x; f_2(x) \geq f_1(x)\}$ is just as good as R_0. [*Hint*: $\alpha + \beta = \sum_R f_1(x) + 1 - \sum_R f_2(x)$.]

1.7 For the bivariate normal probability model in section 1.13, show that when $k = \exp\{(1.96)^2/2\} = 6.82$ the $1/k$ likelihood interval (1.8) is a 95% confidence interval for y. That is, show that the **random interval** defined by replacing x in (1.8) by a random variable which has the conditional probability distribution of X, given $Y = y$, contains the point y with probability 0.95.