4
Elementary hypothesis testing

> I conceive the mind as a moving thing, and arguments as the motive forces driving it in one direction or the other.
>
> *John Craig (1699)*

John Craig was a Scottish mathematician, and one of the first scholars to recognize the merit in Isaac Newton's new invention of 'the calculus'. The above sentence, written some 300 years ago in one of the early attempts to create a mathematical model of reasoning, requires changing by only one word in order to describe our present attitude. We would like to think that our minds are swayed not by arguments, but by evidence. And if fallible humans do not always achieve this objectivity, our desiderata were chosen with the aim of achieving it in our robot. Therefore to see how our robot's mind is 'driven in one direction or the other' by new evidence, we examine some applications that, although simple mathematically, have proved to have practical importance in several different fields.

As is clear from the basic desiderata listed in Chapter 1, the fundamental principle underlying all probabilistic inference is:

> *To form a judgment about the likely truth or falsity of any proposition A, the correct procedure is to calculate the probability that A is true:*
>
> $$P(A|E_1E_2\cdots) \tag{4.1}$$
>
> *conditional on all the evidence at hand.*

In a sampling context (i.e. when A stands for some data set), this principle has seemed obvious to everybody from the start. We used it implicitly throughout Chapter 3 without feeling any need to state it explicitly. But when we turn to a more general context, the principle needs to be stressed because it has not been obvious to all workers (as we shall see repeatedly in later chapters).

The essence of 'honesty' or 'objectivity' demands that we take into account all the evidence we have, not just some arbitrarily chosen subset of it. Any such choice would amount either to ignoring evidence that we have, or presuming evidence that we do not have. This leads us to recognize at the outset that some information is always available to the robot.

4.1 Prior probabilities

Generally, when we give the robot its current problem, we will give it also some new information or 'data' D pertaining to the specific matter at hand. But almost always the robot will have other information which we denote, for the time being, by X. This includes, at the very least, all its past experience, from the time it left the factory to the time it received its current problem. That is always part of the information available, and our desiderata do not allow the robot to ignore it. If we humans threw away what we knew yesterday in reasoning about our problems today, we would be below the level of wild animals; we could never know more than we can learn in one day, and education and civilization would be impossible.

So to our robot there is no such thing as an 'absolute' probability; all probabilities are necessarily conditional on X at least. In solving a problem, its inferences should, according to the principle (4.1), take the form of calculating probabilities of the form $P(A|DX)$. Usually, part of X is irrelevant to the current problem, in which case its presence is unnecessary but harmless; if it is irrelevant, it will cancel out mathematically. Indeed, that is what we really mean by 'irrelevant'.

Any probability $P(A|X)$ that is conditional on X alone is called a *prior probability*. But we caution that the term 'prior' is another of those terms from the distant past that can be inappropriate and misleading today. In the first place, it does not necessarily mean 'earlier in time'. Indeed, the very concept of time is not in our general theory (although we may of course introduce it in a particular problem). The distinction is a purely logical one; any additional information beyond the immediate data D of the current problem is by definition 'prior information'.

For example, it has happened more than once that a scientist has gathered a mass of data, but before getting around to the data analysis he receives some surprising new information that completely changes his ideas of how the data should be analyzed. That surprising new information is, logically, 'prior information' because it is not part of the data. Indeed, the separation of the totality of the evidence into two components called 'data' and 'prior information' is an arbitrary choice made by us, only for our convenience in organizing a chain of inferences. Although all such organizations must lead to the same final results if they succeed at all, some may lead to much easier calculations than others. Therefore, we do need to consider the order in which different pieces of information shall be taken into account in our calculations.

Because of some strange things that have been thought about prior probabilities in the past, we point out also that it would be a big mistake to think of X as standing for some hidden major premise, or some universally valid proposition about Nature. Old misconceptions about the origin, nature, and proper functional use of prior probabilities are still common among those who continue to use the archaic term '*a-priori probabilities*'. The term '*a-priori*' was introduced by Immanuel Kant to denote a proposition whose truth can be known independently of experience; which is most emphatically what we do *not* mean here. X denotes simply whatever additional information the robot has beyond what we have chosen to call 'the data'. Those who are actively familiar with the use of prior probabilities in current real problems usually abbreviate further, and instead of saying 'the prior probability' or 'the prior probability distribution', they say simply, 'the *prior*'.

There is no single universal rule for assigning priors – the conversion of verbal prior information into numerical prior probabilities is an open-ended problem of logical analysis, to which we shall return many times. At present, four fairly general principles are known – group invariance, maximum entropy, marginalization, and coding theory – which have led to successful solutions of many different kinds of problems. Undoubtedly, more principles are waiting to be discovered, which will open up new areas of application.

In conventional sampling theory, the only scenario considered is essentially that of 'drawing from an urn', and the only probabilities that arise are those that presuppose the contents of the 'urn' or the 'population' already known, and seek to predict what 'data' we are likely to get as a result. Problems of this type can become arbitrarily complicated in the details, and there is a highly developed mathematical literature dealing with them. For example, the massive two-volume work of Feller (1950, 1966) and the weighty compendium of Kendall and Stuart (1977) are restricted entirely to the calculation of sampling distributions. These works contain hundreds of nontrivial solutions that are useful in all parts of probability theory, and every worker in the field should be familiar with what is available in them.

However, as noted in the preceding chapter, almost all real problems of scientific inference involve us in the opposite situation; we already know the data D, and want probability theory to help us decide on the likely contents of the 'urn'. Stated more generally, we want probability theory to indicate which of a given set of hypotheses $\{H_1, H_2, \ldots\}$ is most likely to be true in the light of the data and any other evidence at hand. For example, the hypotheses may be various suppositions about the physical mechanism that is generating the data. But fundamentally, as in Chapter 3, physical causation is not an essential ingredient of the problem; what is essential is only that there be some kind of *logical* connection between the hypotheses and the data.

To solve this problem does not require any new principles beyond the product rule (3.1) that we used to find conditional sampling distributions; we need only to make a different choice of the propositions. Let us now use the notation

$$\begin{aligned} X &= \text{prior information,} \\ H &= \text{some hypothesis to be tested,} \\ D &= \text{the data,} \end{aligned}$$

and write the product rule in the form

$$P(DH|X) = P(D|HX)P(H|X) = P(H|DX)P(D|X). \tag{4.2}$$

We recognize $P(D|HX)$ as the sampling distribution which we studied in Chapter 3, but now written in a more flexible notation. In Chapter 3 we did not need to take any particular note of the prior information X, because all probabilities were conditional on H, and so we could suppose implicitly that the general verbal prior information defining the problem was included in H. This is the habit of notation that we have slipped into, which has obscured the unified nature of all inference. Throughout all of sampling theory one can get away with this, and as a result the very term ‘prior information’ is absent from the literature of sampling theory.

Now, however, we are advancing to probabilities that are not conditional on H, but are still conditional on X, so we need separate notations for them. We see from (4.2) that to judge the likely truth of H in the light of the data, we need not only the sampling probability $P(D|HX)$ but also the prior probabilities for D and H:

$$P(H|DX) = P(H|X)\frac{P(D|HX)}{P(D|X)}. \tag{4.3}$$

Although the derivation (4.2)–(4.3) is only the same mathematical result as (3.50)–(3.51), it has appeared to many workers to have a different logical status. From the start it has seemed clear how one determines numerical values of sampling probabilities, but not what determines the prior probabilities. In the present work we shall see that this was only an artifact of an unsymmetrical way of formulating problems, which left them ill-posed. One could see clearly how to assign sampling probabilities because the hypothesis H was stated very specifically; had the prior information X been specified equally well, it would have been equally clear how to assign prior probabilities.

When we look at these problems on a sufficiently fundamental level and realize how careful one must be to specify the prior information before we have a well-posed problem, it becomes evident that there is in fact no logical difference between (3.51) and (4.3); exactly the same principles are needed to assign either sampling probabilities or prior probabilities, and one man’s sampling probability is another man’s prior probability.

The left-hand side of (4.3), $P(H|DX)$, is generally called a ‘*posterior probability*’, with the same *caveat* that this means only ‘logically later in the particular chain of inference being made’, and not necessarily ‘later in time’. And again the distinction is conventional, not fundamental; one man’s prior probability is another man’s posterior probability. There is really only one kind of probability; our different names for them refer only to a particular way of organizing a calculation.

The last factor in (4.3) also needs a name, and it is called the *likelihood* $L(H)$. To explain current usage, we may consider a fixed hypothesis and its implications for different data sets; as we have noted before, the term $P(D|HX)$, in its dependence on D for fixed H, is called the ‘sampling distribution’. But we may consider a fixed data set in the light of various different hypotheses $\{H, H', \ldots\}$; in its dependence on H for fixed D, $P(D|HX)$ is called the ‘likelihood’.

A likelihood $L(H)$ is not itself a probability for H; it is a dimensionless numerical function which, when multiplied by a prior probability and a normalization factor, may become a probability. Because of this, constant factors are irrelevant, and may be struck out. Thus, the quantity $L(H_i) = y(D)\, P(D|H_i X)$ is equally deserving to be called the likelihood, where y is any positive number which may depend on D but is independent of the hypotheses $\{H_i\}$.

Equation (4.3) is then the fundamental principle underlying a wide class of scientific inferences in which we try to draw conclusions from data. Whether we are trying to learn the character of a chemical bond from nuclear magnetic resonance data, the effectiveness of a medicine from clinical data, the structure of the earth's interior from seismic data, the elasticity of a demand from economic data, or the structure of a distant galaxy from telescopic data, (4.3) indicates what probabilities we need to find in order to see what conclusions are justified by the totality of our evidence. If $P(H|DX)$ is very close to one (zero), then we may conclude that H is very likely to be true (false) and act accordingly. But if $P(H|DX)$ is not far from $1/2$, then the robot is warning us that the available evidence is not sufficient to justify any very confident conclusion, and we need to obtain more and better evidence.

4.2 Testing binary hypotheses with binary data

The simplest nontrivial problem of hypothesis testing is the one where we have only two hypotheses to test and only two possible data values. Surprisingly, this turns out to be a realistic and valuable model of many important inference and decision problems. Firstly, let us adapt (4.3) to this binary case. It gives us the probability that H is true, but we could have written it equally well for the probability that H is false:

$$P(\overline{H}|DX) = P(\overline{H}|X)\frac{P(D|\overline{H}X)}{P(D|X)}, \tag{4.4}$$

and if we take the ratio of the two equations,

$$\frac{P(H|DX)}{P(\overline{H}|DX)} = \frac{P(H|X)}{P(\overline{H}|X)}\frac{P(D|H\ X)}{P(D|\overline{H}X)}, \tag{4.5}$$

the term $P(D|X)$ will drop out. This may not look like any particular advantage, but the quantity that we have here, the ratio of the probability that H is true to the probability that it is false, has a technical name. We call it the '*odds*' on the proposition H. So if we write the 'odds on H, given D and X', as the symbol

$$O(H|D\ X) \equiv \frac{P(H|D\ X)}{P(\overline{H}|DX)}, \tag{4.6}$$

then we can combine (4.3) and (4.4) into the following form:

$$O(H|D\ X) = O(H|X)\frac{P(D|HX)}{P(D|\overline{H}X)}. \tag{4.7}$$

The posterior odds on H is (are?) equal to the prior odds multiplied by a dimensionless factor, which is also called a likelihood ratio. The odds are (is?) a strict monotonic function of the probability, so we could equally well calculate this quantity.[1]

[1] Our uncertain phrasing here indicates that 'odds' is a grammatically slippery word. We are inclined to agree with purists who say that it is, like 'mathematics' and 'physics', a singular noun in spite of appearances. Yet the urge to follow the vernacular and treat it as plural is sometimes irresistible, and so we shall be knowingly inconsistent and use it both ways, judging what seems euphonious in each case.

In many applications it is convenient to take the logarithm of the odds because of the fact that we can then add up terms. Now we could take logarithms to any base we please, and this cost the writer some trouble. Our analytical expressions always look neater in terms of natural (base e) logarithms. But back in the 1940s and 1950s when this theory was first developed, we used base 10 logarithms because they were easier to find numerically; the four-figure tables would fit on a single page. Finding a natural logarithm was a tedious process, requiring leafing through enormous old volumes of tables.

Today, thanks to hand calculators, all such tables are obsolete and anyone can find a ten-digit natural logarithm just as easily as a base 10 logarithm. Therefore, we started happily to rewrite this section in terms of the aesthetically prettier natural logarithms. But the result taught us that there is another, even stronger, reason for using base 10 logarithms. Our minds are thoroughly conditioned to the base 10 number system, and base 10 logarithms have an immediate, clear intuitive meaning to all of us. However, we just don't know what to make of a conclusion that is stated in terms of natural logarithms, until it is translated back into base 10 terms. Therefore, we re-rewrote this discussion, reluctantly, back into the old, ugly base 10 convention.

We define a new function, which we will call the *evidence* for H given D and X:

$$e(H|DX) \equiv 10 \log_{10} O(H|DX). \tag{4.8}$$

This is still a monotonic function of the probability. By using the base 10 and putting the factor 10 in front, we are now measuring evidence in *decibels* (hereafter abbreviated to db). The evidence for H, given D, is equal to the prior evidence plus the number of db provided by working out the log likelihood in the last term below:

$$e(H|DX) = e(H|X) + 10 \log_{10} \left[\frac{P(D|HX)}{P(D|\overline{H}X)} \right]. \tag{4.9}$$

Now suppose that this new information D actually consisted of several different propositions:

$$D = D_1 D_2 D_3 \cdots . \tag{4.10}$$

Then we could expand the likelihood ratio by successive applications of the product rule:

$$e(H|DX) = e(H|X) + 10 \log_{10} \left[\frac{P(D_1|H\ X)}{P(D_1|\overline{H}X)} \right] + 10 \log_{10} \left[\frac{P(D_2|D_1\ H\ X)}{P(D_2|D_1\ \overline{H}X)} \right] + \cdots . \tag{4.11}$$

But, in many cases, the probability for getting D_2 is not influenced by knowledge of D_1:

$$P(D_2|D_1 HX) = P(D_2|HX). \tag{4.12}$$

One then says conventionally that D_1 and D_2 are *independent*. Of course, we should really say that the *probabilities which the robot assigns to them* are independent. It is a semantic confusion to attribute the property of 'independence' to propositions or events; for that implies, in common language, physical *causal* independence. We are concerned here with the very different quality of *logical* independence.

To emphasize this, note that neither kind of independence implies the other. Two events may be in fact causally dependent (i.e. one influences the other); but for a scientist who has not yet discovered this, the probabilities representing his state of knowledge – which determine the only inferences he is able to make – might be independent. On the other hand, two events may be causally independent in the sense that neither exerts any causal influence on the other (for example, the apple crop and the peach crop); yet we perceive a logical connection between them, so that new information about one changes our state of knowledge about the other. Then for us their probabilities are not independent.

Quite generally, as the robot's state of knowledge represented by H and X changes, probabilities conditional on them may change from independent to dependent or *vice versa*; yet the real properties of the events remain the same. Then one who attributed the property of dependence or independence to the events would be, in effect, claiming for the robot the power of psychokinesis. We must be vigilant against this confusion between reality and a state of knowledge about reality, which we have called the 'mind projection fallacy'.

The point we are making is not just pedantic nitpicking; we shall see presently (Eq. (4.29)) that it has very real, substantive consequences. In Chapter 3 we have discussed some of the conditions under which these probabilities might be independent, in connection with sampling from a very large known population and sampling with replacement. In the closing Comments section, we noted that whether urn probabilities do or do not factor can depend on whether we do or do not know that the contents of several urns are the same. In our present problem, as in Chapter 3, to interpret causal independence as logical independence, or to interpret logical dependence as causal dependence, has led some to nonsensical conclusions in fields ranging from psychology to quantum theory.

In case these several pieces of data are logically independent given $(H\,X)$ and also given $(\overline{H}X)$, (4.11) becomes

$$e(H|DX) = e(H|X) + 10\sum_i \log_{10}\left[\frac{P(D_i|HX)}{P(D_i|\overline{H}X)}\right], \tag{4.13}$$

where the sum is over all the extra pieces of information that we obtain.

To get some feeling for numerical values here, let us construct Table 4.1. We have three different scales on which we can measure degrees of plausibility: evidence, odds, or probability; they are all monotonic functions of each other. Zero db of evidence corresponds to odds of 1 or to a probability of 1/2. Now, every physicist or electrical engineer knows that 3 db means a factor of 2 (nearly) and 10 db is a factor of 10 (exactly); and so if we go in steps of 3 db, or 10, we can construct this table very easily.

It is obvious from Table 4.1 why it is very cogent to give evidence in decibels. When probabilities approach one or zero, our intuition doesn't work very well. Does the difference between the probability of 0.999 and 0.9999 mean a great deal to you? It certainly doesn't to the writer. But after living with this for only a short while, the difference between evidence of plus 30 db and plus 40 db does have a clear meaning to us. It is now in a scale which our minds comprehend naturally. This is just another example of the Weber–Fechner law; intuitive human sensations tend to be logarithmic functions of the stimulus.

Table 4.1. *Evidence, odds, and probability.*

e	O	p
0	1:1	1/2
3	2:1	2/3
6	4:1	4/5
10	10:1	10/11
20	100:1	100/101
30	1000:1	0.999
40	10^4:1	0.9999
$-e$	$1/O$	$1-p$

Even the factor of 10 in (4.8) is appropriate. In the original acoustical applications, it was introduced so that a 1 db change in sound intensity would be, psychologically, about the smallest change perceptible to our ears. With a little familiarity and a little introspection, we think that the reader will agree that a 1 db change in evidence is about the smallest increment of plausibility that is perceptible to our intuition. Nobody claims that the Weber–Fechner law is a precise rule for all human sensations, but its general usefulness and appropriateness is clear; almost always it is not the absolute change, but more nearly the relative change, in some stimulus that we perceive. For an interesting account of the life and work of Gustav Theodor Fechner (1801–87), see Stigler (1986c).

Now let us apply (4.13) to a specific calculation, which we shall describe as a problem of industrial quality control (although it could be phrased equally well as a problem of cryptography, chemical analysis, interpretation of a physics experiment, judging two economic theories, etc.). Following the example of Good (1950), we assume numbers which are not very realistic in order to elucidate some points of principle. Let the prior information X consist of the following statements:

$X \equiv$ We have 11 automatic machines turning out widgets, which pour out of the machines into 11 boxes. This example corresponds to a very early stage in the development of widgets, because ten of the machines produce one in six defective. The 11th machine is even worse; it makes one in three defective. The output of each machine has been collected in an unlabeled box and stored in the warehouse.

We choose one of the boxes and test a few of the widgets, classifying them as ‘good’ or ‘bad’. Our job is to decide whether we chose a box from the bad machine or not; that is, whether we are going to accept this batch or reject it.

Let us turn this job over to our robot and see how it performs. Firstly, it must find the prior evidence for the various propositions of interest. Let

$$A \equiv \text{we chose a bad batch (1/3 defective)},$$
$$B \equiv \text{we chose a good batch (1/6 defective)}.$$

The qualitative part of our prior information X told us that there are only two possibilities; so in the 'logical environment' generated by X, these propositions are related by negation: given X, we can say that

$$\overline{A} = B, \qquad \overline{B} = A. \tag{4.14}$$

The only quantitative prior information is that there are 11 machines and we do not know which one made our batch, so, by the principle of indifference, $P(A|X) = 1/11$, and

$$e(A|X) = 10 \log_{10} \frac{P(A|X)}{P(\overline{A}|X)} = 10 \log_{10} \frac{(1/11)}{(10/11)} = -10 \text{ db}, \tag{4.15}$$

whereupon we have necessarily $e(B|X) = +10$ db.

Evidently, in this problem the only properties of X that will be relevant for the calculation are just these numbers, ±10 db. Any other kind of prior information which led to the same numbers would give us just the same mathematical problem from this point on. So, it is not necessary to say that we are talking only about a problem where there are 11 machines, and so on. There might be only one machine, and the prior information consists of our previous experience with it.

Our reason for stating the problem in terms of 11 machines was that we have, thus far, only one principle, indifference, by which we can convert raw information into numerical probability assignments. We interject this remark because of a famous statement by Feller (1950) about a single machine, which we consider in Chapter 17 after accumulating some more evidence pertaining to the issue he raised. To our robot, it makes no difference how many machines there are; the only thing that matters is the prior probability for a bad batch, however this information was arrived at.[2]

Now, from this box we take out a widget and test it to see whether it is defective. If we pull out a bad one, what will that do to the evidence for a bad batch? That will add to it

$$10 \log_{10} \frac{P(\text{bad}|A\ X)}{P(\text{bad}|\overline{A}X)} \text{ db} \tag{4.16}$$

where $P(\text{bad}|AX)$ represents the probability for getting a bad widget, given A, etc.; these are sampling probabilities, and we have already seen how to calculate them. Our procedure is very much 'like' drawing from an urn, and, as in Chapter 3, on one draw our datum D now consists only of a binary choice: (good/bad). The sampling distribution $P(D|HX)$

[2] Notice that in this observation we have the answer to a point raised in Chapter 1: How does one make the robot 'cognizant' of the semantic meanings of the various propositions that it is being called upon to deal with? The answer is that the robot does not need to be 'cognizant' of anything. If we give it, in addition to the model and the data, a list of the propositions to be considered, with their prior probabilities, this conveys all the 'meaning' needed to define the robot's mathematical problem for the applications now being considered. Later, we shall wish to design a more sophisticated robot which can also help us to assign prior probabilities by analysis of complicated but incomplete information, by the maximum entropy principle. But, even then, we can always define the robot's mathematical problem without going into semantics.

reduces to

$$P(\text{bad}|AX) = \frac{1}{3}, \qquad P(\text{good}|AX) = \frac{2}{3}, \tag{4.17}$$

$$P(\text{bad}|BX) = \frac{1}{6}, \qquad P(\text{good}|BX) = \frac{5}{6}. \tag{4.18}$$

Thus, if we find a bad widget on the first draw, this will increase the evidence for A by

$$10 \log_{10} \frac{(1/3)}{(1/6)} = 10 \log_{10} 2 = 3 \text{ db}. \tag{4.19}$$

What happens now if we draw a second bad one? We are sampling without replacement, so as we noted in (3.11), the factor (1/3) in (4.19) should be updated to

$$\frac{(N/3) - 1}{N - 1} = \frac{1}{3} - \frac{2}{3(N-1)}, \tag{4.20}$$

where N is the number of widgets in the batch. But, to avoid this complication, we suppose that N is very much larger than any number that we contemplate testing; i.e. we are going to test such a negligible fraction of the batch that the proportion of bad and good ones in it is not changed appreciably by the drawing. Then the limiting form of the hypergeometric distribution (3.22) will apply, namely the binomial distribution (3.86). Thus we shall consider that, given A or B, the probability for drawing a bad widget is the same at every draw regardless of what has been drawn previously; so every bad one we draw will provide $+3$ db of evidence in favor of hypothesis A.

Now suppose we find a good widget. Using (4.14), we get evidence for A of

$$10 \log_{10} \frac{P(\text{good}|AX)}{P(\text{good}|BX)} = 10 \log_{10} \frac{(2/3)}{(5/6)} = -0.97 \text{ db}, \tag{4.21}$$

but let's call it -1 db. Again, this will hold for any draw, if the number in the batch is sufficiently large. If we have inspected n widgets, of which we found n_b bad ones and n_g good ones, the evidence that we have the bad batch will be

$$e(A|DX) = e(A|X) + 3n_\text{b} - n_\text{g}. \tag{4.22}$$

You see how easy this is to do once we have set up the logarithmic machinery. The robot's mind is 'driven in one direction or the other' in a very simple, direct way.

Perhaps this result gives us a deeper insight into why the Weber–Fechner law applies to intuitive plausible inference. Our 'evidence' function is related to the data that we have observed in about the most natural way imaginable; a given increment of evidence corresponds to a given increment of data. For example, if the first 12 widgets we test yield five bad ones, then

$$e(A|DX) = -10 + 3 \times 5 - 7 = -2 \text{ db}, \tag{4.23}$$

or, the probability for a bad batch is raised by the data from $(1/11) = 0.09$ to $P(A|DX) \simeq 0.4$.

In order to get at least 20 db of evidence for proposition A, how many bad widgets would we have to find in a certain sequence of $n = n_\text{b} + n_\text{g}$ tests? This requires

$$3n_\text{b} - n_\text{g} = 4n_\text{b} - n = n(4f_\text{b} - 1) \geq 20, \tag{4.24}$$

so, if the fraction $f_\text{b} \equiv n_\text{b}/n$ of bad ones remains greater than 1/4, we shall accumulate eventually 20 db, or any other positive amount, of evidence for A. It appears that $f_\text{b} = 1/4$ is the threshold value at which the test can provide no evidence for either A or B over the other; but note that the $+3$ and -1 in (4.22) are only approximate. The exact threshold fraction of bad ones is, from (4.19) and (4.21),

$$f_\text{t} = \frac{\log\left(\frac{5}{4}\right)}{\log(2) + \log\left(\frac{5}{4}\right)} = 0.2435292, \tag{4.25}$$

in which the base of the logarithms does not matter. Sampling fractions greater (less) than this give evidence for A over B (B over A); but if the observed fraction is close to the threshold, it will require many tests to accumulate enough evidence.

Now all we have here is the probability or odds or evidence, whatever you wish to call it, of the proposition that we chose the bad batch. Eventually, we have to make a decision: we're going to accept it, or we're going to reject it. How are we going to do that? Well, we might decide beforehand: if the probability of proposition A reaches a certain level, then we'll decide that A is true. If it gets down to a certain value, then we'll decide that A is false.

There is nothing in probability theory *per se* which can tell us where to put these critical levels at which we make our decision. This has to be based on value judgments: what are the consequences of making wrong decisions, and what are the costs of making further tests? This takes us into the realm of decision theory, considered in Chapters 13 and 14. But for now it is clear that making one kind of error (accepting a bad batch) might be more serious than making the other kind of error (rejecting a good batch). That would have an obvious effect on where we place our critical levels.

So we could give the robot some instructions such as 'If the evidence for A is greater than $+0$ db, then reject this batch (it is more likely to be bad than good). If it goes as low as -13 db, then accept it (there is at least a 95% probability that it is good). Otherwise, continue testing.' We start doing the tests, and every time we find a bad widget the evidence for the bad batch goes up 3 db; every time we find a good one, it goes down 1 db. The tests terminate as soon as we enter either the accept or reject region for the first time.

The way described above is how our robot would do it if we told it to reject or accept on the basis that the *posterior probability* of proposition A reaches a certain level. This very useful and powerful procedure is called 'sequential inference' in the statistical literature, the term signifying that the number of tests is not determined in advance, but depends on the sequence of data values that we find; at each step in the sequence we make one of three choices: (a) stop with acceptance; (b) stop with rejection; (c) make another test. The term should not be confused with what has come to be called 'sequential analysis with nonoptional stopping', which is a serious misapplication of probability theory; see the discussions of optional stopping in Chapters 6 and 17.

4.3 Nonextensibility beyond the binary case

The binary hypothesis testing problem turned out to have such a beautifully simple solution that we might like to extend it to the case of more than two hypotheses. Unfortunately, the convenient independent additivity over data sets in (4.13) and the linearity in (4.22) do not generalize. By 'independent additivity' we mean that the increment of evidence from a given datum D_i depends only on D_i and H; not on what other data have been observed. As (4.11) shows, we always have additivity, but not independent additivity unless the probabilities are independent.

We state the reason for this nonextensibility in the form of an exercise for the reader; to prepare for it, suppose that we have n hypotheses $\{H_1, \ldots, H_n\}$ which on prior information X are mutually exclusive and exhaustive:

$$P(H_i H_j|X) = P(H_i|X)\delta_{ij}, \qquad \sum_{i=1}^{n} P(H_i|X) = 1. \tag{4.26}$$

Also, we have acquired m data sets $\{D_1, \ldots, D_m\}$, and as a result the probabilities of the H_i become updated in odds form by (4.7) , which now becomes

$$O(H_i|D_1, \ldots, D_m X) = O(H_i|X)\frac{P(D_1, \ldots, D_m|H_i X)}{P(D_1, \ldots, D_m|\overline{H}_i X)}. \tag{4.27}$$

It is common that the numerator will factor because of the logical independence of the D_j, given H_i:

$$P(D_1, \ldots, D_m|H_i X) = \prod_j P(D_j|H_i X), \quad 1 \le i \le n. \tag{4.28}$$

If the denominator should also factor,

$$P(D_1, \ldots, D_m|\overline{H}_i X) = \prod_j P(D_j|\overline{H}_i X), \quad 1 \le i \le n, \tag{4.29}$$

then (4.27) would split into a product of the updates produced by each D_j separately, and the log-odds formula (4.9) would again take a form independently additive over the D_j as in (4.13).

Exercise 4.1. Show that there is no such nontrivial extension of the binary case. More specifically, prove that if (4.28) and (4.29) hold with $n > 2$, then at most one of the factors

$$\frac{P(D_1|H_i\ X)}{P(D_1|\overline{H}_i\ X)} \cdots \frac{P(D_m|H_i\ X)}{P(D_m|\overline{H}_i\ X)} \tag{4.30}$$

is different from unity, therefore at most one of the data sets D_j can produce any updating of the probability for H_i.

This has been a controversial issue in the literature of artificial intelligence (Glymour, 1985; R. W. Johnson, 1985). Those who fail to distinguish between logical independence and causal independence would suppose that (4.29) is always valid, provided only that no D_i exerts a physical influence on any other D_j. But we have already noted the folly of such reasoning; this is an occasion when the semantic confusion can lead to serious numerical errors. When $n = 2$, (4.29) follows from (4.28). But when $n > 2$, (4.29) is such a strong condition that it would reduce the whole problem to a triviality not worth considering; we have left it (Exercise 4.1) for the reader to examine the equations to see why this is so. Because of Cox's theorems expounded in Chapter 2, the verdict of probability theory is that our conclusion about nonextensibility can be evaded only at the price of committing demonstrable inconsistencies in our reasoning.

To head off a possible misunderstanding of what is being said here, let us add the following. However many hypotheses we have in mind, it is of course always possible to pick out two of them and compare them only against each other. This reverts to the binary choice case already analyzed, and the independent additive property holds within that smaller problem (find the status of an hypothesis relative to a single alternative).

We could organize this by choosing A_1 as the standard 'null hypothesis' and comparing each of the others with it by solving $n - 1$ binary problems; whereupon the relative status of any two propositions is determined. For example, if A_5 and A_7 are favored over A_1 by 22.3 db and 31.9 db, respectively, then A_7 is favored over A_5 by $31.9 - 22.3 = 9.6$ db. If such binary comparisons provide all the information one wants, there is no need to consider multiple hypothesis testing at all.

But that would not solve our present problem; given the solutions of all these binary problems, it would still require a calculation as big as the one we are about to do to convert that information into the absolute status of any given hypothesis relative to the entire class of n hypotheses. Here we are going after the solution of the larger problem directly.

In any event, we need not base our stance merely on claims of authoritarian finality for an abstract theorem; more constructively, we now show that probability theory does lead us to a definite, useful procedure for multiple hypothesis testing, which gives us a much deeper insight and makes it clear why the independent additivity cannot, *and should not*, hold when $n > 2$. It would then ignore some very cogent information; that is the demonstrable inconsistency.

4.4 Multiple hypothesis testing

Suppose that something very remarkable happens in the sequential test just discussed: we tested 50 widgets and every one turned out to be bad. According to (4.22), that would give us 150 db of evidence for the proposition that we had the bad batch. $e(A|E)$ would end up at $+140$ db, which is a probability which differs from unity by one part in 10^{14}. Now, our common sense rejects this conclusion; some kind of innate skepticism rises in us. If you test 50 widgets and you find that all 50 are bad, you are not willing to believe that you have a batch in which only one in three are really bad. So what went wrong here? Why doesn't our robot work in this case?

We have to recognize that our robot is immature; it reasons like a four-year-old child does. The remarkable thing about small children is that you can tell them the most ridiculous things and they will accept it all with wide open eyes, open mouth, and it never occurs to them to question you. They will believe anything you tell them.

Adults learn to make mental allowance for the reliability of the source when told something hard to believe. One might think that, ideally, the information which our robot should have put into its memory was not that we had either 1/3 bad or 1/6 bad; the information it should have put in was that some unreliable human *said* that we had either 1/3 bad or 1/6 bad.

More generally, it might be useful in many problems if the robot could take into account the fact that the information it has been given may not be perfectly reliable to begin with. There is always a small chance that the prior information or data that we fed to the robot was wrong. In a real problem there are always hundreds of possibilities, and if you start out the robot with dogmatic initial statements which say that there are only two possibilities, then of course you must not expect its conclusions to make sense in every case.

To accomplish this skeptically mature behavior automatically in a robot is something that we can do, when we come to consider significance tests; but fortunately, after further reflection, we realize that for most problems the present immature robot is what we want after all, because we have better control over it.

We *do* want the robot to believe whatever we tell it; it would be dangerous to have a robot who suddenly became skeptical in a way not under our control when we tried to tell it some true but startling – and therefore highly important – new fact. But then the onus is on us to be aware of this situation, and when there is a good chance that skepticism will be needed, it is up to us to give the robot a hint about how to be skeptical for that particular problem.

In the present problem we can give the hint which makes the robot skeptical about A when it sees 'too many' bad widgets, by providing it with one more possible hypothesis, which notes that possibility and therefore, in effect, puts the robot on the lookout for it. As before, let proposition A mean that we have a box with 1/3 defective, and proposition B is the statement that we have a box with 1/6 bad. We add a third proposition, C, that something went entirely wrong with the machine that made our widgets, and it is turning out 99% defective.

Now we have to adjust our prior probabilities to take this new possibility into account. But we do not want this to be a major change in the nature of the problem; so let hypothesis C have a very low prior probability $P(C|X)$ of 10^{-6} (-60 db). We could write out X as a verbal statement which would imply this, but as in the previous footnote we can state what proposition X is, with no ambiguity at all for the robot's purposes, simply by giving it the probabilities conditional on X, of all the propositions that we're going to use in this problem. In that way we don't state everything about X that is important to us conceptually; but we state everything about X that is relevant to the robot's current mathematical problem.

So, suppose we start out with these initial probabilities:

$$\begin{aligned} P(A|X) &= \frac{1}{11}(1 - 10^{-6}), \\ P(B|X) &= \frac{10}{11}(1 - 10^{-6}), \\ P(C|X) &= 10^{-6}, \end{aligned} \tag{4.31}$$

where

$$\begin{aligned} A &\equiv \text{we have a box with 1/3 defective,} \\ B &\equiv \text{we have a box with 1/6 defective,} \\ C &\equiv \text{we have a box with 99/100 defective.} \end{aligned}$$

The factors $(1 - 10^{-6})$ are practically negligible, and for all practical purposes we will start out with the initial values of evidence:

$$\begin{aligned} &-10 \text{ db for } A, \\ &+10 \text{ db for } B, \\ &-60 \text{ db for } C. \end{aligned} \tag{4.32}$$

The data proposition D stands for the statement that 'm widgets were tested and every one was defective'. Now, from (4.9), the posterior evidence for proposition C is equal to the prior evidence plus ten times the logarithm of this probability ratio:

$$e(C|DX) = e(C|X) + 10 \log_{10} \frac{P(D|CX)}{P(D|\overline{C}X)}. \tag{4.33}$$

Our discussion of sampling with and without replacement in Chapter 3 shows that

$$P(D|CX) = \left(\frac{99}{100}\right)^m \tag{4.34}$$

is the probability that the first m are all bad, given that 99% of the machine's output is bad, under our assumption that the total number in the box is large compared with the number m tested.

We also need the probability $P(D|\overline{C}X)$, which we can evaluate by two applications of the product rule (4.3):

$$P(D|\overline{C}X) = P(D|X) \frac{P(\overline{C}|DX)}{P(\overline{C}|X)}. \tag{4.35}$$

In this problem, the prior information states dogmatically that there are only three possibilities, and so the statement $\overline{C} \equiv$ 'C is false' implies that either A or B must be true:

$$P(\overline{C}|DX) = P(A + B|DX) = P(A|DX) + P(B|DX), \tag{4.36}$$

where we used the general sum rule (2.66), the negative term dropping out because A and B are mutually exclusive. Similarly,

$$P(\overline{C}|X) = P(A|X) + P(B|X). \tag{4.37}$$

Now, if we substitute (4.36) into (4.35), the product rule will be applicable again in the form

$$\begin{aligned} P(AD|X) &= P(D|X)P(A|DX) = P(A|X)P(D|AX) \\ P(BD|X) &= P(D|X)P(B|DX) = P(B|X)P(D|BX), \end{aligned} \tag{4.38}$$

and so (4.35) becomes

$$P(D|\overline{C}X) = \frac{P(D|AX)P(A|X) + P(D|BX)P(B|X)}{P(A|X) + P(B|X)}, \tag{4.39}$$

in which all probabilities are known from the statement of the problem.

4.4.1 Digression on another derivation

Although we have the desired result (4.39), let us note that there is another way of deriving it, which is often easier than direct application of (4.3). The principle was introduced in our derivation of (3.33): resolve the proposition whose probability is desired (in this case D) into mutually exclusive propositions, and calculate the sum of their probabilities. We can carry out this resolution in many different ways by ‘introducing into the conversation’ any set of mutually exclusive and exhaustive propositions $\{P, Q, R, \dots\}$ and using the rules of Boolean algebra:

$$D = D(P + Q + R + \cdots) = DP + DQ + DR + \cdots. \tag{4.40}$$

But the success of the method depends on our cleverness at choosing a particular set for which we can complete the calculation. This means that the propositions introduced must have a known kind of relevance to the question being asked; the example of penguins at the end of Chapter 2 will not be helpful if that question has nothing to do with penguins.

In the present case, for evaluation of $P(D|\overline{C}X)$, it appears that propositions A and B have this kind of relevance. Again, we note that proposition $\overline{C}$ implies $(A + B)$; and so

$$\begin{aligned} P(D|\overline{C}X) &= P(D(A + B)|\overline{C}X) = P(DA + DB|\overline{C}X) \\ &= P(DA|\overline{C}X) + P(DB|\overline{C}X). \end{aligned} \tag{4.41}$$

These probabilities can be factored by the product rule:

$$P(D|\overline{C}X) = P(D|A\overline{C}X)P(A|\overline{C}X) + P(D|B\overline{C}X)P(B|\overline{C}X). \tag{4.42}$$

But we can abbreviate: $P(D|A\overline{C}X) \equiv P(D|AX)$ and $P(D|B\overline{C}X) \equiv P(D|BX)$, because, in the way we set up this problem, the statement that either A or B is true implies that C must be false. For this same reason, $P(\overline{C}|AX) = 1$, and so, by the product rule,

$$P(A|\overline{C}X) = \frac{P(A|X)}{P(\overline{C}|X)}, \tag{4.43}$$

and similarly for $P(B|\overline{C}X)$. Substituting these results into (4.42) and using (4.37), we again arrive at (4.39). This agreement provides another illustration – and a rather severe test – of the consistency of our rules for extended logic.

Returning to (4.39), we have the numerical value

$$P(D|\overline{C}X) = \left(\frac{1}{3}\right)^m \left(\frac{1}{11}\right) + \left(\frac{1}{6}\right)^m \frac{10}{11}, \tag{4.44}$$

and everything in (4.33) is now at hand. If we put all these things together, we find that the evidence for proposition C is:

$$e(C|DX) = -60 + 10\log_{10}\left[\frac{\left(\frac{99}{100}\right)^m}{\frac{1}{11}\left(\frac{1}{3}\right)^m + \frac{10}{11}\left(\frac{1}{6}\right)^m}\right]. \tag{4.45}$$

If $m > 5$, a good approximation is

$$e(C|DX) \simeq -49.6 + 4.73\, m, \qquad m > 5, \tag{4.46}$$

and if $m < 3$, a crude approximation is

$$e(C|DX) \simeq -60 + 7.73\, m, \qquad m < 3. \tag{4.47}$$

Proposition C starts out at -60 db, and the first few bad widgets we find will each give about 7.73 db of evidence in favor of C, so the graph of $e(C|DX)$ vs. m will start upward at a slope of 7.73. But then the slope drops, when $m > 5$, to 4.73. The evidence for C reaches 0 db when $m \simeq 49.6/4.73 = 10.5$. So, ten consecutive bad widgets would be enough to raise this initially very improbable hypothesis by 58 db, to the place where the robot is ready to consider it very seriously; and 11 consecutive bad ones would take it over the threshold, to where the robot considers it more likely to be true than false.

In the meantime, what is happening to our propositions A and B? As before, A starts off at -10 db, B starts off at $+10$ db, and the plausibility for A starts going up 3 db per defective widget. But after we've found too many bad ones, that skepticism would set in, and you and I would begin to doubt whether the evidence really supports proposition A after all; proposition C is becoming a much easier way to explain what is observed. Has the robot also learned to be skeptical?

After m widgets have been tested, and all proved to be bad, the evidence for propositions A and B, and the approximate forms, are as follows:

$$\begin{aligned} e(A|DX) &= -10 + 10\log_{10}\left[\frac{\left(\frac{1}{3}\right)^m}{\left(\frac{1}{6}\right)^m + \frac{11}{10}\times 10^{-6}\left(\frac{99}{100}\right)^m}\right] \\ &\simeq \begin{Bmatrix} -10 + 3m & \text{for } m < 7 \\ +49.6 - 4.73m & \text{for } m > 8 \end{Bmatrix} \end{aligned} \tag{4.48}$$

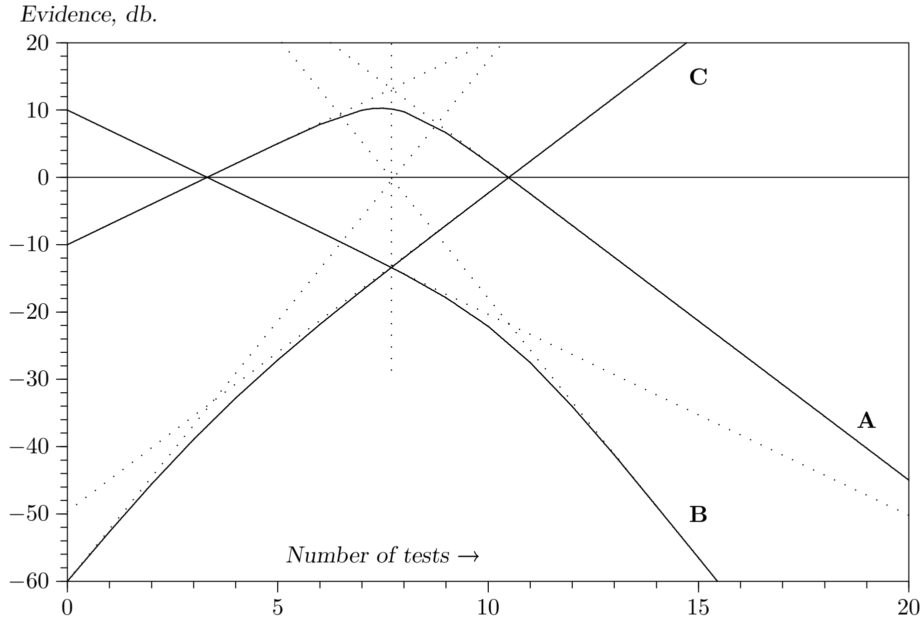

Fig. 4.1. A surprising multiple sequential test wherein a dead hypothesis (C) is resurrected.

$$e(B|DX) = +10 + 10\log_{10}\left[\frac{\left(\frac{1}{6}\right)^m}{\left(\frac{1}{3}\right)^m + 11\times 10^{-6}\left(\frac{99}{100}\right)^m}\right]$$
$$\simeq \left\{\begin{array}{ll} 10 - 3m & \text{for } m < 10 \\ 59.6 - 7.33m & \text{for } m > 11 \end{array}\right\}. \tag{4.49}$$

The exact results are summarized in Figure 4.1. We can learn quite a lot about multiple hypothesis testing from studying this diagram. The initial straight line part of the A and B curves represents the solution as we found it before we introduced proposition C; the change in plausibility for propositions A and B starts off just the same as in the previous problem. The effect of proposition C does not appear until we have reached the place where C crosses B. At this point, suddenly the character of the A curve changes; instead of going on up, at $m = 7$ it has reached its highest value of 10 db. Then it turns around and comes back down; the robot has indeed learned how to become skeptical. But the B curve does *not* change at this point; it continues on linearly until it reaches the place where A and C have the same plausibility, and at this point it has a change in slope. From then on, it falls off more rapidly.

Most people find all this surprising and mysterious at first glance; but then a little meditation is enough to make us perceive what is happening and why. The change in plausibility for A due to one more test arises from the fact that we are now testing hypothesis A against two alternatives: B and C. But, initially, B is so much more plausible than C, that for all practical purposes we are simply testing A against B, and reproducing our previous solution (4.22). After enough evidence has accumulated to bring the plausibility for C up to the same level as B, then from that point on A is essentially being tested against C instead of B, which is a very different situation.

All of these changes in slope can be interpreted in this way. Once we see this principle, it is clear that the same thing is going to be true more generally. As long as we have a discrete set of hypotheses, a change in plausibility for any one of them will be approximately the result of a test of this hypothesis against a single alternative – the single alternative being that one of the remaining hypotheses which is most plausible at that time. As the relative plausibilities of the alternatives change, the slope of the A curve must also change; *this is the cogent information that would be lost* if we tried to retain the independent additive form (4.13) when $n > 2$.

Whenever the hypotheses are separated by about 10 db or more, then multiple hypothesis testing reduces approximately to testing each hypothesis against a single alternative. So, seeing this, you can construct curves of the sort shown in Fig. 4.1 very rapidly without even writing down the equations, because what would happen in the two-hypothesis case is easily seen once and for all. The diagram has a number of other interesting geometrical properties, suggested by drawing the six asymptotes and noting their vertical alignment (dotted lines), which we leave for the reader to explore.

All the information needed to construct fairly accurate charts resulting from any sequence of good and bad tests is contained in the 'plausibility flow diagrams' of Figure 4.2, which summarize the solutions of all those binary problems; every possible way to test one proposition against a single alternative. It indicates, for example, that finding a good widget raises the evidence for B by 1 db if B is being tested against A, and by 19.22 db if it is being tested against C. Similarly, finding a bad widget raises the evidence for A by 3 db if A is being tested against B, but lowers it by 4.73 db if it is being tested against C. Likewise, we see that finding a single good widget lowers the evidence for C by an amount that cannot be recovered by two bad ones; so there is a 'threshold of skepticism'. C will never attain an appreciable probability; i.e. the robot will never become skeptical about propositions A and B, as long as the observed fraction f of bad ones remains less than 2/3.

More precisely, we define a threshold fraction f_{t} thus: as the number of tests $m \to \infty$ with $f = m_{\mathrm{b}}/m \to$ const., $e(C|DX)$ tends to $+\infty$ if $f > f_{\mathrm{t}}$, and to $-\infty$ if $f < f_{\mathrm{t}}$. The exact threshold turns out to be greater than 2/3: $f_{\mathrm{t}} = 0.793951$ (Exercise 4.2). If the observed

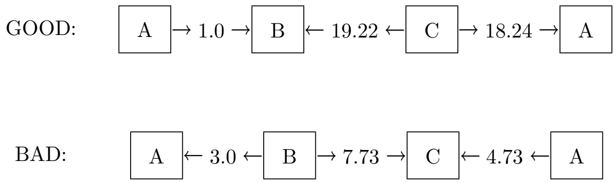

Fig. 4.2. Plausibility flow diagrams.

fraction of bad widgets remains above this value, the robot will be led eventually to prefer proposition C over A and B.

Exercise 4.2. Calculate the exact threshold of skepticism $f_t(x, y)$, supposing that proposition C has instead of 10^{-6} an arbitrary prior probability $P(C|X) = x$, and specifies instead of 99/100 an arbitrary fraction y of bad widgets. Then discuss how the dependence on x and y corresponds – or fails to correspond – to human common sense. *Hint:* In problems like this, always try first to get an analytic solution in closed form. If you are unable to do this, then you must write a short computer program which will display the correct numerical values in tables or graphs.

Exercise 4.3. Show how to make the robot skeptical about both unexpectedly high and unexpectedly low numbers of bad widgets in the observed sample. Give the full equations. Note particularly the following: if A is true, then we would expect, according to the binomial distribution (3.86), that the observed fraction of bad ones would tend to about 1/3 with many tests, while if B is true it should tend to 1/6. Suppose that it is found to tend to the threshold value (4.24), close to 1/4. On sufficiently large m, you and I would then become skeptical about A and B; but intuition tells us that this would require a much larger m than ten, which was enough to make us and the robot skeptical when we find them all bad. Do the equations agree with our intuition here, if a new hypothesis F is introduced which specifies $P(\text{bad}|FX) \simeq 1/4$?

In summary, the role of our new hypothesis C was only to be held in abeyance until needed, like a fire extinguisher. In a normal testing situation it is ‘dead’, playing no part in the inference because its probability is and remains far below that of the other hypotheses. But a dead hypothesis can be resurrected to life by very unexpected data. Exercises 4.2 and 4.3 ask the reader to explore the phenomenon of resurrection of dead hypotheses in more detail than we do in this chapter, but we return to the subject in Chapter 5.

Figure 4.1 shows an interesting thing. Suppose we had decided to stop the test and accept hypothesis A if the evidence for it reached +6 db. As we see, it would overshoot that value at the sixth trial. If we stopped the testing at that point, then we would never see the rest of this curve and see that it really goes down again. If we had continued the testing beyond this point, then we would have changed our minds again.

At first glance this seems disconcerting, but notice that it is inherent in all problems of hypothesis testing. If we stop the test at any finite number of trials, then we can never be absolutely sure that we have made the right decision. It is always possible that still more tests would have led us to change our decision. But note also that probability theory as logic has automatic built-in safety devices that can protect us against unpleasant surprises. Although it is always *possible* that our decision is wrong, this is extremely *improbable* if our critical level for decision requires $e(A|DX)$ to be large and positive. For example, if $e(A|DX) \geq 20$ db, then $P(A|DX) > 0.99$, and the total probability for all the alternatives is less than 0.01; then few of us would hesitate to decide confidently in favor of A.

In a real problem we may not have enough data to give such good evidence, and we might suppose that we could decide safely if the most likely hypothesis A is well separated from the alternatives, even though $e(A|DX)$ is itself not large. Indeed, if there are 1000 alternatives but the separation of A from the most likely alternative is more than 20 db, then the odds favor A by more than 100:1 over any one of the alternatives, and if we were obliged to make a definite choice of one hypothesis here and now, there could still be no hesitation in choosing A; it is clearly the best we can do with the information we have. Yet we cannot do it so confidently, for it is now very plausible that the decision is wrong, because the class of alternatives as a whole is about as probable as A. But probability theory warns us, by the numerical value of $e(A|DX)$, that this is the case; we need not be surprised by it.

In scientific inference our job is always to do the best we can with whatever information we have; there is no advance guarantee that our information will be sufficient to lead us to the truth. But many of the supposed difficulties arise from an inexperienced user's failure to recognize and use the safety devices that probability theory as logic always provides. Unfortunately, the current literature offers little help here because its viewpoint, concentrated mainly on sampling theory, directs attention to other things such as assumed sampling frequencies, as the following exercises illustrate.

Exercise 4.4. Suppose that B is in fact true; estimate how many tests it will probably require in order to accumulate an additional 20 db of evidence (above the prior 10 db) in favor of B. Show that the sampling probability that we could ever obtain 20 db of evidence for A is negligibly small, even if we sample millions of times. In other words it is, for all practical purposes, impossible for a doctrinaire zealot to sample to a foregone false conclusion merely by continuing until he finally gets the evidence he wants.
Note: The calculations called for here are called 'random walk' problems; they are sampling theory exercises. Of course, the results are not wrong, only incomplete. Some essential aspects of inference in the real world are not recognized by sampling theory.

Exercise 4.5. The estimate asked for in Exercise 4.4 is called the 'average sample number' (ASN), and the original rationale for the sequential procedure (Wald, 1947) was not our derivation from probability theory as logic, but Wald's conjecture (unproven at the time) that the sequential probability-ratio tests such as (4.19) and (4.21) minimize the ASN for a given reliability of conclusion. Discuss the validity of this conjecture; can one define the term 'reliability of conclusion' in such a way that the conjecture can be proved true?

Evidently, we could extend this example in many different directions. Introducing more 'discrete' hypotheses would be perfectly straightforward, as we have seen. More interesting would be the introduction of a continuous range of hypotheses, such as

$$H_f \equiv \text{the machine is putting out a fraction } f \text{ bad.}$$

Then, instead of a discrete prior probability distribution, our robot would have a continuous distribution in $0 \le f \le 1$, and it would calculate the posterior probabilities for various values of f on the basis of the observed samples, from which various decisions could be made. In fact, although we have not yet given a formal discussion of continuous probability distributions, the extension is so easy that we can give it as an introduction to this example.

4.5 Continuous probability distribution functions

Our rules for inference were derived in Chapter 2 only for the case of finite sets of discrete propositions $(A, B, \ldots)$. But this is all we ever need in practice. Suppose that f is any continuously variable real parameter of interest, then the propositions

$$\begin{aligned} F' &\equiv (f \le q) \\ F'' &\equiv (f > q) \end{aligned} \tag{4.50}$$

are discrete, mutually exclusive, and exhaustive; so our rules will surely apply to them. Given some information Y, the probability for F' will in general depend on q, defining a function

$$G(q) \equiv P(F'|Y), \tag{4.51}$$

which is evidently monotonic increasing. Then what is the probability that f lies in any specified interval $(a < f \le b)$? The answer is probably obvious intuitively, but it is worth noting that it is determined uniquely by the sum rule of probability theory, as follows. Define the propositions

$$A \equiv (f \le a), \qquad B \equiv (f \le b), \qquad W \equiv (a < f \le b). \tag{4.52}$$

Then a relation of Boolean algebra is $B = A + W$, and since A and W are mutually exclusive, the sum rule reduces to

$$P(B|Y) = P(A|Y) + P(W|Y). \tag{4.53}$$

But $P(B|Y) = G(b)$, and $P(A|Y) = G(a)$, so we have the result

$$P(a < f \le b|Y) = P(W|Y) = G(b) - G(a). \tag{4.54}$$

In the present case, $G(q)$ is continuous and differentiable, so we may write also

$$P(a < f \le b|Y) = \int_a^b \mathrm{d}f g(f), \tag{4.55}$$