CHAPTER 4

Bayesian Versus Non-Bayesian Approaches

In this chapter we shall consider how, by attributing positive probabilities to hypotheses in the manner described in Chapter 2, one can account for many of the characteristic features of scientific practice, particularly as they relate to deterministic theories.

a THE BAYESIAN NOTION OF CONFIRMATION

Information gathered in the course of observation is often considered to have a bearing on the acceptability of a theory or hypothesis (we use the terms interchangeably), either by confirming it or by disconfirming it. Such information may either derive from casual observation or, more commonly, from experiments deliberately contrived in the hope of obtaining relevant evidence. The idea that evidence may count for or against a theory, or be neutral towards it, is a central feature of scientific inference, and the Bayesian account will clearly need to start with a suitable interpretation of these concepts.

Fortunately, there is a suitable and very natural interpretation, for if $P(h)$ measures your belief in a hypothesis when you do not know the evidence e, and $P(h \mid e)$ is the corresponding measure when you do, e surely confirms h when the latter exceeds the former. So we shall take the following as our definitions:

e **confirms or supports** h when $P(h \mid e) > P(h)$

e **disconfirms or undermines** h when $P(h \mid e) < P(h)$

e **is neutral with respect to** h when $P(h \mid e) = P(h)$

One might reasonably take $P(h \mid e) - P(h)$ as measuring the degree of e's support for h, though other measures have been suggested (e.g., Good, 1950). Disagreements on this score will not be controversial in this book. We shall refer, in the usual way, to $P(h)$ as 'the prior probability of h' and to $P(h \mid e)$ as h's 'posterior probability' relative to, or in the light of, e. The reasons for this terminology are obvious, but it ought to be noted that the terms have a meaning only in relation to evidence: as Lindley (1970, p. 38) put it, "Today's posterior distribution is tomorrow's prior". It should be remembered too that all the probabilities are evaluated in relation to accepted background knowledge.

■ b THE APPLICATION OF BAYES'S THEOREM

Bayes's Theorem relates the posterior probability of a hypothesis, $P(h \mid e)$, to the terms $P(h)$, $P(e \mid h)$, and $P(e)$. Hence, knowing the values of these last three terms, it is possible to determine whether e confirms h, and, more importantly, to calculate $P(h \mid e)$. In practice, of course, the various probabilities may only be known rather imprecisely; we shall have more to say about this practical aspect of the question later.

The dependence of the posterior probability on the three terms referred to above is reflected in three striking phenomena of scientific inference. First, other things being equal, the extent to which evidence e confirms a hypothesis h increases with the likelihood of h on e, that is to say, with $P(e \mid h)$. At one extreme, where e refutes h, $P(e \mid h) = 0$; hence, disconfirmation is at a maximum. The greatest confirmation is produced, for a given $P(e)$, when $P(e \mid h) = 1$, which will be met in practice when h logically entails e. Statistical hypotheses, which will be dealt with in parts III and IV of this book, are more substantially confirmed the higher the value of $P(e \mid h)$.

Secondly, the posterior probability of a hypothesis depends on its prior probability, a dependence sometimes discernible in scientific attitudes to ad hoc hypotheses and in frequently expressed preferences for the simpler of two hypotheses. As we shall see, scientists always discriminate, in advance of any experimentation, between theories they regard as more or less credible and, so, worthy of attention and others.

Thirdly, the power of e to confirm h depends on $P(e)$, that is to say, on the probability of e when it is not assumed that h is true (which, of course, is not the same as assuming h to be false). This dependence is reflected in the scientific intuition that

the more surprising the evidence, the greater its confirming power. However, $P(e) = P(e \mid h)P(h) + P(e \mid {\sim}h)P({\sim}h)$, as we showed in Chapter 2, section **e**, so that really, the posterior probability of h depends on the three basic quantities $P(h)$, $P(e \mid h)$, and $P(e \mid {\sim}h)$.

We shall deal in greater detail with each of these facets of inductive reasoning in the course of this chapter.

■ c FALSIFYING HYPOTHESES

A characteristic pattern of scientific inference is the refutation of a theory, when one of the theory's empirical consequences has been shown to be false in an experiment. As we saw, this kind of reasoning, with its straightforward and unimpeachable logical structure, exercised such an influence on Popper that he made it into the centrepiece of his scientific philosophy.

Although the Bayesian approach was not conceived specifically with this aspect of scientific reasoning in view, it has a ready explanation for it. The explanation relies on the fact that if, relative to background knowledge, a hypothesis h entails a consequence e, then (relative to the same background knowledge) $P(h \mid {\sim}e) = 0$. Interpreted in the Bayesian fashion, this means that h is maximally disconfirmed when it is refuted. Moreover, it can be shown that, as we should expect, once a theory is refuted, no further evidence can confirm it, unless the evidence or some part of the background assumptions are revoked. (This is simply proved: if h entails e, then $h \,\&\, {\sim}e$ is a contradiction, so $P(h \,\&\, {\sim}e) = 0$, whence $P(h \mid e) = 0$. And if f is some further datum, then since $h \,\&\, {\sim}e \,\&\, f$ is also a contradiction, the same argument shows that $P(h \mid {\sim}e \,\&\, f) = 0$.)

■ d CHECKING A CONSEQUENCE

A standard method of investigating a deterministic hypothesis is to draw out some of its logical consequences, relative to some stock of background theories, and check whether they are true or not. For instance, the General Theory of Relativity was confirmed by establishing that light is deflected when it passes near the sun, as the theory predicts. It is easy to show, by means of Bayes's Theorem, why and under what circumstances a theory is confirmed by its consequences.

If h entails e, then, as may be simply shown, $P(e \mid h) = 1$. Hence, from Bayes's Theorem: $P(h \mid e) = \frac{P(h)}{P(e)}$. Thus, if $0 < P(e) < 1$, and if $P(h) > 0$, then $P(h \mid e) > P(h)$. It follows that any evidence whose probability is neither of the extreme values must confirm every hypothesis with a non-zero probability of which it is a logical consequence.

Succeeding confirmations must eventually diminish in force, for the theory has an upper limit of probability, beyond which no amount of evidence can push it. And as the theory becomes more probable with the accumulation of evidence, further consequences of the theory acquire a greater likelihood of being true, and thus a smaller power to confirm. All this follows from Bayes's Theorem. Suppose $e_1, e_2, \ldots e_n \ldots$ are consequences of h, which are found to be true. Then Bayes's Theorem asserts that

$$P(h \mid e_1 \,\&\, e_2 \ldots \&\, e_n) = \frac{P(h)}{P(e_1 \,\&\, e_2 \ldots \&\, e_n)}$$

Now

$$P(e_1 \,\&\, e_2 \ldots \&\, e_n) = P(e_1)P(e_2 \,\&\, \ldots \&\, e_n \mid e_1)$$

and

$$P(e_2 \,\&\, \ldots \&\, e_n \mid e_1) = P(e_2 \mid e_1)P(e_3 \,\&\, \ldots \&\, e_n \mid e_1 \,\&\, e_2)$$

Thus, in general,

$$P(e_1 \,\&\, e_2 \,\&\, \ldots \&\, e_n) = P(e_1)P(e_2 \mid e_1) \ldots P(e_n \mid e_1 \,\&\, \ldots \&\, e_{n-1})$$

Hence,

$$P(h \mid e_1 \,\&\, e_2 \,\&\, \ldots \&\, e_n) = \frac{P(h)}{P(e_1)P(e_2 \mid e_1) \ldots P(e_n \mid e_1 \,\&\, \ldots \&\, e_{n-1})}.$$

Provided $P(h) > 0$, the term $P(e_n \mid e_1 \,\&\, \ldots \&\, e_{n-1})$ must tend to 1. If it did not, the posterior probability of h would at some point exceed 1, which is impossible (Jeffreys, 1961, pp. 43–44). This explains why one would not continue to test a hypothesis indefinitely, though without more detailed information on the individual's belief-structure, in particular regarding the values of $P(e_n \mid e_1 \,\&\, \ldots \&\, e_{n-1})$, one could not predict the precise point beyond which further predictions of the hypothesis were sufficiently probable not to be worth examining.

Specific categories of a theory's consequences also have a restricted capacity to confirm (Urbach, 1981). Suppose h is the theory under discussion and that h_r is a substantial restriction of that theory. A substantial restriction of Newton's theory might, for example, express the idea that freely falling bodies near the earth descend with a constant acceleration or that the period and length of a pendulum are related by the familiar formula. Since h entails h_r, $P(h) \leq P(h_r)$ (*see* Chapter 2, section **e**), and if h_r is much less speculative than its progenitor, it will often be significantly more probable.

Now consider a series of predictions derived from h, but which also follow from h_r. These may then confirm both theories, their posterior probabilities being given by Bayes's Theorem, thus:

$$P(h \mid e_1 \,\&\, e_2 \ldots \&\, e_n) = \frac{P(h)}{P(e_1 \,\&\, e_2 \ldots \&\, e_n)}$$

and

$$P(h_r \mid e_1 \,\&\, e_2 \ldots \&\, e_n) = \frac{P(h_r)}{P(e_1 \,\&\, e_2 \ldots \&\, e_n)}.$$

Combining these two equations to eliminate the common denominator, one obtains

$$P(h \mid e_1 \,\&\, e_2 \ldots \&\, e_n) = \frac{P(h)}{P(h_r)} \times P(h_r \mid e_1 \,\&\, e_2 \ldots \&\, e_n).$$

Since the maximum value of the last probability term in this equation is 1, it follows that however many predictions of h_r are verified, the main theory, h, can never acquire a posterior probability in excess of $\frac{P(h)}{P(h_r)}$. Hence, the type of evidence characterised by entailment from h_r may well be limited in its capacity to confirm h. This explains the phenomenon that repetitions of an experiment often confirm a general theory only to a limited extent, for the predictions verified by means of a given kind of experiment (that is, an experiment designed to a specified pattern) do normally follow from and confirm a much restricted version of the predicting theory.

When an experiment's capacity to generate confirming evidence has been exhausted through repetition, further support would have to be sought from other experiments, moreover, experiments of *different kinds*. We have an intuitive grasp on the idea of diversity among experiments. For instance, measuring the melting point of oxygen on a Monday and on a Tuesday would be the same experiment, but would be different from determining the rate at which oxygen and hydrogen react to form water. Ascertaining this reaction rate under different temperature and pressure conditions would presumably also count as different experiments, though it seems natural to say in such cases that the differences are not so great.

Franklin and Howson (1984) characterised similarity amongst experiments in a way which does considerable justice to these intuitions. They considered two experiments, E and E', each capable, in principle, of being instantiated indefinitely and yielding, respectively, the outcomes $e_1, e_2, \ldots$ and $e'_1, e'_2, \ldots$. They suggested that E and E' are different just in case for all $m > m_o$, and for some m_o,

$$P(e_{m+1} \mid e_1 \,\&\, e_2 \ldots \&\, e_m) > P(e'_i \mid e_1 \,\&\, e_2 \ldots \&\, e_m)$$

and for all $n > n_o$, and for some n_o,

$$P(e'_{n+1} \mid e'_1 \,\&\, e'_2 \ldots \&\, e'_n) > P(e_j \mid e'_1 \,\&\, e'_2 \ldots \&\, e'_n).$$

What this condition states, in other words, is that beyond a certain number of repetitions of E (respectively E'), the probability of a further outcome of that experiment is greater than the probability of any outcome of E' (respectively E). When this is the case, Franklin and Howson say that E and E' are different. (There is clearly scope for extending this definition to cover the notion of degrees of difference between experiments.) The definition means that if all the experimental outcomes follow deductively from some hypothesis, and if one experiment had been performed sufficiently often, then that hypothesis would be more substantially confirmed by the outcome of a different experiment than by a further instance of the same one, which is what we set out to show. The closely related fact that a wide variety of data gives greater support to a hypothesis than an equally extensive collection of similar data will be discussed later in the chapter, under a different heading (section **j.5**).

The arguments and explanations in this section rely on the possibility that evidence already accumulated from an experiment may increase the probability of further performances of that experiment producing similar results. Such a possibility is contested by Popperians, who rule it out as an unacceptable deviation from a purely deductive pattern of reasoning. But in so doing, they appear to rule out any explanation for the fact, attested by every scientist, that by repeating some experiment, one eventually (usually quickly) exhausts its capacity to confirm a given hypothesis. Alan Musgrave (1975), however, thought the fact could be explained non-inductively, in a manner compatible with Popperian principles. He claimed that after a certain number of repetitions of an experiment, the scientist would form a generalisation to the effect that whenever the experiment is performed, it yields a similar result. Musgrave then suggested that the generalisation would be entered into 'background knowledge'. Relative to this newly augmented background knowledge, the experiment is certain to produce a similar result on its next performance. Musgrave then appealed to the principle that evidence confirms a hypothesis in proportion to the difference between its probability relative to the hypothesis together with background knowledge and its probability relative to background knowledge alone. (That is, in Popper's notation, confirmation is proportional to $P(e \mid h \,\&\, b) - P(e \mid b)$, where b is background knowledge.) Musgrave then inferred that even if the experiment did produce the expected result when next performed, the hypothesis would receive no new confirmation. Watkins (1984, p. 297) more recently concurred with this account.

A number of objections may be made against it, though. First, as we shall show in the next section, although it seems to be a fact and is an essential constituent of Bayesian reasoning, there is no basis in the Popperian methodology for confirmation to depend on the probability of the evidence; Popper simply invoked the principle ad hoc. Secondly, Musgrave's suggestion takes no account of the fact that a given experimental result may be generalised in infinitely many ways. This is a substantial objection since, clearly, different generalisations give rise to different expectations about the outcomes of future experiments. Musgrave's account is incomplete without some rule to specify in each case the appropriate generalisation that should be formulated and adopted. Finally, the decision to designate the generalisation background knowledge, with the consequent effect on our evaluation of other theories and on our future conduct regarding, for example, whether to repeat certain experiments, is comprehensible only if we have invested some confidence in the theory. But then Musgrave's account tacitly calls on the same kind of inductive considerations as it was designed to circumvent, so its aim is defeated.

e THE PROBABILITY OF THE EVIDENCE

The degree to which h is confirmed by e depends, according to Bayesian theory, on the extent to which $P(e \mid h)$ exceeds $P(e)$, that is, on how much more probable e is relative to the hypothesis and background assumptions than it is relative just to background assumptions. Another way of putting this is to say that confirmation is correlated with how much more probable the evidence is if the hypothesis is true than if it is false. This is obvious from Bayes's Theorem when it is reformulated as follows:

$$\frac{P(h \mid e)}{P(h)} = \frac{P(e \mid h)}{P(e)} = \frac{1}{P(h) + \dfrac{P(e \mid \sim h)}{P(e \mid h)} P(\sim h)} .$$

These facts are reflected in the everyday experience that information that is particularly unexpected or surprising unless some hypothesis is assumed to be true, supports that hypothesis with particular force. Thus, if a soothsayer predicts that you will meet a dark stranger sometime and you do in fact, your faith in his powers of precognition would not be much enhanced: you would probably continue to think his predictions were just the result of guesswork. However, if the prediction also gave the correct number of hairs on the head of that stranger, your previous scepticism would no doubt be severely shaken.

Cox (1961, p. 92) illustrated this point with an incident in *Macbeth*. The three witches, using their special brand of divination, predicted to Macbeth that he would soon become both Thane of Cawdor and King of Scotland. He finds both these prognostications almost impossible to believe:

> By Sinel's death, I know I am Thane of Glamis,
> But how of Cawdor?
> The Thane of Cawdor lives, a prosperous gentleman,
> And to be King stands not within the prospect of belief,
> No more than to be Cawdor.

But a short time later he learns that the Thane of Cawdor prospered no longer and was in fact dead and that he, Macbeth, has succeeded to the title. As a result, Macbeth's attitude to the witches' powers is entirely altered and he comes to believe in their other predictions and in their ability to foresee the future.

The following, more scientific, example was used by Jevons (1874, vol. 1, pp. 278–279) to illustrate the dependence of confirmation on the improbability of the evidence. The distinguished scientist, Charles Babbage, examined numerous logarithmic tables published over two centuries in various parts of the world. He was interested in whether they were derived from the same source or had been worked out independently. Babbage (1827) found the same six errors in all but two and drew the "irrestistible" conclusion that, apart from these two, all the tables originated in a common source.

Babbage's reasoning was interpreted by Jevons roughly as follows. The theory, t_1, which says of some pair of logarithmic tables that they had a common origin, is moderately likely, in view of the immense amount of labour needed to compile such tables ab initio, and for a number of other reasons. The alternative independence theory might take a variety of forms, each attributing different probabilities to the occurrence of errors in various positions in the table. The only one of these which seems at all likely would assign each place an equal probability of exhibiting an error and would, moreover, regard these errors as being more or less independent. Call this theory t_2 and let e^i be the evidence of i common errors in the tables. The posterior probability of t_1 is inversely proportional to $P(e^i)$, which, under the assumption of only two rival hypotheses, can be expressed as $P(e^i) = P(e^i \mid t_1)P(t_1) + P(e^i \mid t_2)P(t_2)$. (This is the theorem of total probability—*see* Chapter 2, section **e.**) Since t_1 entails e^i, $P(e^i) = P(t_1) + P(e^i \mid t_2)P(t_2)$. The quantity $P(e^i \mid t_2)$ clearly decreases with increasing i. Hence $P(e^i)$ diminishes and tends to $P(t_1)$, as i increases; and so e^i becomes increasingly powerful evidence for t_1, a result which agrees with scientific intuition.

In fact, scientists seem to regard a few shared mistakes in different mathematical tables as so strongly indicative of a common source that at least one compiler of such tables attempted to protect his copyright by deliberately incorporating three minor errors "as a trap for would-be plagiarists" (L. J. Comrie, quoted by Bowden, 1953, p. 4).

The relationship between how surprising a piece of evidence is on background assumptions and its power to confirm a hypothesis is a natural consequence of the Bayesian theory and was not deliberately built in. On the other hand, approaches that eschew probabilistic assessments of hypotheses and attempt to base scientific method on deductive logic alone seem constitutionally incapable of accounting for the phenomenon. Such approaches would need to be able, first, to discriminate between items of evidence on grounds other than their deductive or probabilistic relation to a hypothesis. And having established such a basis for discriminating, they must show a connection with confirmation. The objectivist school has more or less dodged this challenge. An exception is Popper. In tackling the problem, he moved partway towards Bayesianism; however, the concessions he made were insufficient. Thus Popper conceded that, in regard to confirmation, the significant quantities are $P(e \mid h)$ and $P(e)$, and he even measured the degree to which e confirms h (or "corroborates" it, to use Popper's preferred term) by the difference between these quantities. (Popper, 1959a, appendix *ix)

But Popper never stated explicitly what he meant by the probability of evidence. On the one hand, he would never have allowed it to have a subjective connotation, for that would have compromised the supposed objectivity of science; on the other hand, he never worked out what objective significance the term could have. His writings suggest that he had in mind some purely logical notion of probability, but as we saw in Chapter 3, there is no adequate account of logical probability. Popper also never explained satisfactorily why a hypothesis benefits from improbable evidence or, to put the objection another way, he failed to provide a foundation in non-Bayesian terms for the Bayesian confirmation function which he appropriated. (For a discussion and decisive criticism of Popper's account, see Grünbaum, 1976.)

The Bayesian position has recently been misunderstood to imply that if some evidence is known, then it cannot support any hypothesis, on the grounds that known evidence must have unit probability. That the objection is based on a misunderstanding is explained in Chapter 11, where a number of other criticisms of the Bayesian approach will be rebutted.

■ f THE RAVENS PARADOX

That evidence supports a hypothesis more the greater the ratio $\frac{P(e \mid h)}{P(e)}$ scotches a famous puzzle first posed by Hempel (1945)

and known as the *Paradox of Confirmation* or sometimes as the *Ravens Paradox*. It was called a paradox because its premisses were regarded as extremely plausible, despite their counterintuitive, or in some versions contradictory, implications, and the reference to ravens stems from the paradigm hypothesis ('All ravens are black') which is frequently used to expound the problem. The difficulty arises from three assumptions about confirmation. They are as follows:

1. Hypotheses of the form 'All R's are B' are confirmed by the evidence of something that is both R and B. For example, 'All ravens are black' is confirmed by a black raven. (Hempel called this Nicod's condition, after the philosopher, Jean Nicod.)
2. Logically equivalent hypotheses are confirmed by the same evidence. (This is the Equivalence condition.)
3. Evidence of some object not being R does not confirm 'All R's are B'.

We shall describe an object that is both black and a raven with the term RB. Similarly, a non-black, non-raven will be denoted $\overline{R}\overline{B}$. A contradiction arises for the following reasons: RB confirms 'All R's are B', on account of the Nicod condition. According to the Equivalence condition, it also confirms 'All non-B's are non-R's', since the two hypotheses are logically equivalent. But contradicting this, the third condition implies that RB does not confirm 'All non-B's are non-R's'.

The contradiction may be avoided by revoking the third condition. (We shall note later another reason for not holding on to it.) However, although the remaining conditions are compatible, they have a consequence which many philosophers have regarded as blatantly false, namely that a non-black, non-raven (say, a red herring or a white shoe) can confirm the hypothesis that all ravens are black. (The argument is this: 'All non-B's are non-R' is equivalent to 'All R's are B'; according to the Nicod condition, the first is confirmed by $\overline{R}\overline{B}$; hence, by the Equivalence condition, so is the second.)

If non-black, non-ravens support the raven hypothesis, this seems to imply the paradoxical result that one could investigate that and other generalisations of a similar form *just as well* by observing white paper and red ink from the comfort of one's writing desk as by studying ravens on the wing. However, this would be a non sequitur. For the fact that RB and $\overline{R}\overline{B}$ both confirm a hypothesis does not imply that they do so with equal force. Once it is recognised that confirmation is a matter of degree, the conclusion is no longer so counterintuitive, because it is compatible with $\bar{R}\bar{B}$ confirming 'All *R*'s are *B*', but to a minuscule and negligible degree.

In fact, this is what Bayesians have maintained. In the particular case of the hypothesis of the ravens, Mackie (1963) argued that since non-black, non-ravens form such a numerous class compared with black ravens, it is almost (but not absolutely) certain that a random object about which we know nothing will turn out to be neither black nor a raven, but relatively unlikely that it will be a black raven. Hence, for a Bayesian, both kinds of object confirm 'All ravens are black', but non-black, non-ravens do so only minutely.

Although the Nicod and Equivalence conditions are not undermined by their implication that the raven hypothesis is confirmed by non-ravens, there are nevertheless good reasons for rejecting the Nicod condition. (The Equivalence condition seems incontestable.) As Good (1961) first demonstrated, 'All *R*'s are *B*' is not necessarily confirmed by an *RB* and, contrary to Nicod, could even be disconfirmed by such an instance. Consider the following example of this effect, which we have taken, with some modification, from Swinburne (1971): 'All grasshoppers are located outside the county of Yorkshire'. The observation of a grasshopper just beyond the county border is an instance of this generalisation and, according to Nicod, confirms it. But it might be more reasonably argued that since there are no border controls restricting the movement of grasshoppers, the observation of one on the edge of the county increases the probability that others have actually entered, and hence undermines the hypothesis. In Bayesian terms, this is a case where the probability of some datum is reduced by a hypothesis (that is, $P(e \mid h) < P(e)$) which is therefore disconfirmed (in other words, $P(h \mid e) < P(h)$).

The grasshopper example also provides an instance where a datum of the type $\bar{R}B$ confirms a generalisation of the form 'All *R*'s are *B*'. Imagine that an object which looks for all the world like a grasshopper were found hopping about just outside Yorkshire and that it turned out to be some other sort of insect. The discovery that the object was not a grasshopper would be relatively unlikely unless the grasshopper hypothesis were true (hence, $P(e) < P(e \mid h)$); thus it would confirm that hypothesis. If the deceptively grasshopper-like object were within the county boundary, the same conclusion would follow, though the

degree of confirmation would be greater. This shows that 'All *R*'s are *B*' may also be confirmed by a datum of the type $\bar{R}\bar{B}$. Hence, the impression that non-*R*'s never confirm such hypotheses may be dispelled.

It is sometimes maintained that 'All ravens are black' would be differently confirmed if a known raven were revealed to be black than if an object were first observed to be black and later found to be a raven. For instance, Horwich, who denoted the first datum R^*B and the second RB^*, argued that the former is the more powerfully confirming instance, on the alleged grounds that only it subjects the hypothesis to the risk of falsification, for the raven could have turned out to be non-black, in which case the hypothesis would have been refuted. By contrast, Horwich said, the latter does not jeopardise the hypothesis, for the black object is compatible with the hypothesis whether it is a raven or not.

This argument, however, is specious. The observation of an object that enquiry reveals to be a black raven poses absolutely no risk of refutation to the hypothesis, however the enquiry was conducted. The only difference between R^*B and RB^* is in the point at which one learns that the hypothesis has not been refuted. This does not seem to us a sufficient reason to distinguish the two data from the point of view of their confirming power. To do so would appear to depart from normal practice, for scientists do not as a rule attach any importance to the distinction. (For a fuller discussion of this point, the reader is referred to Chapter 11, section **g.**)

Our conclusions are, first, that the supposedly paradoxical consequences of Nicod's condition and the Equivalence condition are not problematic, and, secondly, that there are separate reasons for rejecting Nicod's condition, which, moreover, conform to Bayesian principles.

■ g THE DESIGN OF EXPERIMENTS

Not every experiment is equally worth doing and because of the expense that experiments often necessitate, both in labour and in equipment, careful attention is frequently devoted to their design, in order to ensure that they will yield information economically.

What is a well-designed experiment? The natural answer is that it is an experiment which stands a good chance of producing a decisive or, at least, an almost decisive result. The experiment should be decisive in the sense that one hypothesis becomes certainly true or at least almost certainly true, or that as many as possible of the initially most plausible hypotheses become (almost) certainly not true. This allows for the possibility that a poorly constructed experiment may, unexpectedly, produce decisive evidence, while a well-designed experiment may yield an outcome which is quite indecisive.

The considerations that are pertinent to the design of efficient experiments can be appreciated by referring to Bayes's Theorem. Suppose n rival hypotheses, $h_1, \ldots, h_n$, are being entertained and that these are regarded as the only serious contenders for the truth, in the sense that their total probability is 1 or close to 1. We are, as we said, interested in acquiring decisive evidence, that is, the kind of evidence, e, which makes $P(h_i \mid e)$ approach 1 for some h_i or which brings as many as possible of the terms $P(h_j \mid e)$ close to 0. Consider now an experiment, one of whose possible outcomes, e, would have the effect of massively confirming or disconfirming one of the hypotheses. Such evidence would be decisive in our sense. Clearly, the larger $P(e)$, the greater the probability of achieving a decisive result and, hence, the better the experiment.

but sure this is, [illegible] [illegible] model [illegible]

However, a slightly odd fact emerges at this point. In order to confirm a hypothesis strongly, one requires evidence e for which $P(e)$ is low, relative to $P(e \mid h)$. On the other hand, in order for the experiment to be worth doing at all, $P(e)$ should be moderately high. Therefore, two separate considerations determine how well designed an experiment is, and these frequently pull in opposite directions.

When deciding which experiment to perform, one must also take at least three other factors into account: the cost of the experiment; the morality of carrying it out; and the value, both theoretical and practical, of the hypotheses one is interested in. Bayes's Theorem, of course, implies nothing about how these separate factors should be balanced.

h THE DUHEM PROBLEM

h.1 The Problem

The so-called Duhem (or Duhem-Quine) problem is a problem for theories of science of the type associated with Popper, which

emphasise the power of certain evidence to refute a hypothesis. According to Popper's influential views, the characteristic of a theory which makes it 'scientific' is its falsifiability: "Statements or systems of statements, in order to be ranked as scientific, must be capable of conflicting with possible, or conceivable, observations" (Popper, 1963, p. 39). And, claiming to apply this criterion, Popper (1963, ch. 1) judged Einstein's gravitational theory to be scientific and Freud's psychology, unscientific.

There is a strong flavour of commendation about the term scientific which has proved extremely misleading. For a theory which is scientific in Popper's sense is not necessarily true, or even probably true or so much as close to the truth, nor can it be said definitely that it is likely to lead to the truth. In fact, there seems to be no conceptual connection between a theory's capacity to pass Popper's test of scientificness and its having any epistemic or inductive value. There is little alternative, then, so far as we can see, to regarding Popper's demarcation between scientific and unscientific statements as part of a theory about the content and character of what is usually termed science, not as having any normative significance.

Yet as a contribution to understanding the methods of science, Popper's ideas bear little fruit. His central claim was that scientific theories are falsifiable by "possible, or conceivable, observations". This poses a difficulty, for an observation can only falsify a theory (that is, conclusively demonstrate its falsity) if it is itself conclusively certain. But observations cannot be conclusively certain. For instance, the statement 'The hand on this dial is pointing to the numeral 6' is clearly fallible—it is unlikely, but possible, that the person reporting it missaw the position of the hand. The same is true of introspective perceptual reports, such as 'In my visual field there is now a silvery crescent against a dark blue background'. It has recently been maintained (Watkins, 1984, pp. 79 and 248) that this and similar statements "may rightly be regarded by their authors when they make them as infallibly true". But this is not so, for it is possible, though not probable, that the introspector has misremembered and mistaken the shape he usually describes as a crescent or the sensation he usually receives on reporting a blue image. These and other sources of error ensure that introspective reports are not exempt from the rule that non-analytic statements are fallible.

Can you do better? Jeffrey conditioning?

Of course, the kinds of observation statement we have mentioned, if asserted under appropriate circumstances, would never be seriously doubted. That is, although they could be false, they have a force and immediacy that carries conviction; they are 'morally certain', to use the traditional phrase. But if observation statements are merely indubitable, then whether a theory is regarded as refuted by observational data or not rests ultimately on a subjective feeling of certainty. The fact that such convictions are so strong and uncontroversial may disguise their fallibility, but cannot undo it. Hence, no theory is falsifiable, for none could be conclusively shown to be false by empirical observations. In practice the closest one could get to a refutation would be arriving at the conclusion that a theory that clashes with almost certainly true observations is almost certainly false.

Remember this
When it
Comes to
Convergence

Forgets
on p 290

A second objection to Popper's falsifiability criterion, and the one upon which we shall focus for its more general interest, is that it describes as unscientific most of those theories which are usually deemed science's greatest achievements. This is the chief aspect of the well-known criticisms advanced by Polanyi, Kuhn, and Lakatos, amongst others. They have pointed out that, as had already been established by Duhem (1905), many notable theories of science are not falsifiable by what would generally be regarded as observation statements, even if those statements were infallibly true. Predictions drawn from Newton's laws or from the kinetic theory of gases turn out to depend not only on those theories but also on certain auxiliary theories. Hence, if such predictions fail, one is not compelled by logic to infer that the main theory is false, for the fault may lie with one or more of the auxiliary assumptions. The history of science has many occasions when an important theory led to a false prediction and where that theory, nevertheless, was not blamed for the failure. In such cases we find that one or more of the auxiliary assumptions used to derive the prediction was taken to be the culprit. The problem that arose from Duhem's investigations was which of the several distinct theories involved in deriving a false prediction should be regarded as the false element or elements in the assumptions.

h.2 Lakatos's and Kuhn's Treatment of the Duhem Problem

Lakatos examined in detail the way that scientists react to anomalies; indeed, he made it a central feature of what he referred to as his "methodology of scientific research programmes". Lakatos claimed that scientific research of the most significant kind usually proceeds in what he called "research programmes". A research programme takes the form of a central or "hard core" theory, together with an associated "protective belt" of auxiliary assumptions. The function of the latter is to combine with the hard core in order to draw out specific predictions, which can then be checked by experiment. The auxiliary assumptions are described as protective because during a research programme's lifetime they, not the central theory, are revised if a prediction is shown to be false.

Lakatos suggested Newtonian physics as an example of a research programme, the three laws of mechanics and the law of gravitation constituting the hard core, while various optical theories, assumptions about the number and positions of the planets, and so forth, he included in the protective belt. He also described a set of heuristic rules by which the research programme dealt with anomalies and advanced into new areas.

Kuhn's famous theory of scientific paradigms is similar to the methodology we have just described and was probably its inspiration. Both Lakatos and Kuhn were impressed that scientists tend to give the benefit of the doubt to some, especially fundamental, theories when these encounter anomalies—and both argued that such theories exerted a commanding influence over whole areas of scientific research. Lakatos's methodology has the advantage in that it describes scientific research programmes in some detail and that it analyses their modes of action; whereas Kuhn left his corresponding notion of a paradigm somewhat vague in comparison.

Lakatos also outlined criteria of success for a research programme. He held that it was perfectly legitimate to systematically treat the hard core as the innocent party in a refutation, provided the research programme occasionally leads to successful novel predictions or to successful or "non–ad hoc" explanations of existing data. Lakatos called such programmes "progressive".

> The sophisticated falsificationist [which Lakatos counted himself as] . . . sees nothing wrong with a group of brilliant scientists conspiring to pack everything they can into their favourite research programme ('conceptual framework', if you wish) with a sacred hard core. As long as their genius—and luck—en-

> ables them to expand their programme *'progressively'*, while sticking to its hard core, they are allowed to do it. (Lakatos, 1970, p. 187)

If, on the other hand, the programme persistently produced false predictions, or if its explanations were habitually ad hoc, Lakatos called it "degenerating". (We shall devote the next section to the notion of ad hocness.) Lakatos employed these tendentious terms even though he never succeeded in substantiating their intimations of approval and disapproval, and in the end he seems to have abandoned the attempt and settled on the more modest claim that, as a matter of historical fact, progressive programmes have usually been well regarded by scientists, while degenerating ones were distrusted and eventually dropped.

This last claim has, it seems to us, some truth to it, as evidenced, for example, by the case studies in the history of science included in Howson (1976). But although Lakatos and Kuhn identified and described an important aspect of scientific work, they provided no rationale or explanation for it. For instance, Lakatos was never able to explain why a research programme's occasional predictive or explanatory success could compensate for numerous failures, nor could he specify how many such successes are needed to convert a degenerating programme into a progressive one. (They should occur "now and then", he said.) Hence, although the methodology of scientific research programmes points to some of the factors relevant to scientific change, it provides no explanation.

Lakatos was also unable to explain why some theories are raised to the status of the hard core of a research programme and are defended by a protective belt of hypotheses, while others are left to their own devices. From Lakatos's writings, one could think that the question is decided by the scientist's mere whim (Lakatos called it a "methodological fiat"). Unfortunately, this suggests that it is a perfectly canonical scientific practice to set up any theory whatever as the hard core of a research programme, or as the central pattern of a paradigm, and to blame all empirical difficulties on auxiliary theories. This is far from being the case.

h.3 The Duhem Problem Solved by Bayesian Means

The questions left unanswered by Lakatos are answered with the help of Bayes's Theorem, as Dorling (1979) has shown. First we shall consider how the probabilities of several theories are altered when, as a group, they have been refuted.

Suppose a theory, *t*, and an auxiliary hypothesis, *a*, together imply an empirical consequence, which is shown to be false by the observation of the outcome *e*. Let us assume that while the combination *t* & *a* is refuted by *e*, the two components taken separately are not refuted. We wish to consider the separate effects wrought on the probabilities of *t* and *a* by the adverse evidence *e*. The comparisons of interest here are between $P(t \mid e)$ and $P(t)$ and between $P(a \mid e)$ and $P(a)$. The conditional probabilities can be expressed using Bayes's Theorem, as follows:

$$P(t \mid e) = \frac{P(e \mid t)P(t)}{P(e)} \qquad P(a \mid e) = \frac{P(e \mid a)P(a)}{P(e)}.$$

In order to evaluate the posterior probabilities of *t* and of *a*, one must first determine the values of the various terms on the right-hand sides of these equations. Before doing this, it is worth noting that these expressions convey no expectation that the refutation of *t* & *a* jointly considered will in general have a symmetrical effect on the separate probabilities of *t* and of *a*, nor any reason why the degree of asymmetry may not be very large in some cases. Also, the expressions allow one to discern the factors that determine which hypothesis suffers most in the refutation. In particular, the probability of *t* changes very little if $P(e \mid t) \approx P(e)$, while that of *a* is reduced substantially just in case $P(e \mid a)$ is substantially less than $P(e)$.

A historical example might best illustrate how a theory that produces a false prediction may still remain very probable; we shall, in fact, use an example that Lakatos (1970, pp. 138–140, and 1968, pp. 174–175) drew heavily on. In 1815, William Prout, a medical practitioner and chemist, advanced the hypothesis that the atomic weights of all the elements are whole-number multiples of the atomic weight of hydrogen, the underlying assumption being that all matter is built out of different combinations of some basic element. Prout believed hydrogen to be that fundamental building-block, though the idea was entertained by others that a more basic element might exist, out of which hydrogen itself was composed. Now the atomic weights recorded at the time, though close to being integers when expressed as multiples of the atomic weight of hydrogen, did not match Prout's hypothesis exactly. However, these deviations from a perfect fit failed to convince Prout that his hypothesis was wrong; he instead took the view that there were faults in the methods that had been used to measure the relative weights of atoms. Thomas Thomson drew a similar conclusion. Indeed, both he and Prout went so far as to adjust several reported atomic weights in order to bring them into line with Prout's hypothesis. For instance, instead of accepting 0.829 as the atomic weight (expressed as a proportion of the weight of an atom of oxygen) of the element boron, which was the experimentally reported value, Thomson (1818, p. 340) preferred 0.875 "because it is a multiple of 0.125, which all the atoms seem to be". (Thomson erroneously took 0.125 as the atomic weight of hydrogen, relative to that of oxygen.) Similarly, Prout adjusted the measured atomic weight of chlorine, which (relative to hydrogen) was 35.83, to 36.

Thomson's and Prout's reasoning can be explained as follows: Prout's hypothesis t, together with an appropriate assumption a asserting the accuracy (within specified limits) of the measuring technique, the purity of the chemicals employed, and so forth, implies that the measured atomic weight of chlorine (relative to hydrogen) is a whole number. Suppose, as was the case in 1815, that chlorine's measured atomic weight was 35.83, and call this the evidence e. It seems that chemists of the early nineteenth century, such as Prout and Thomson, were fairly certain about the truth of t, but less so of a, though more sure that a is true than that it is false. Contemporary near-certainty about the truth of Prout's hypothesis is witnessed by the chemist J. S. Stas. He reported (1860, p. 42) that "In England the hypothesis of Dr Prout was almost universally accepted as absolute truth", and he confessed that when he started researching into the matter, he himself had "had an almost absolute confidence in the exactness of Prout's principle" (1860, p. 44). (Stas's confidence eventually faded after many years' experimental study, and by 1860 he had "reached the complete conviction, the entire certainty, as far as certainty can be attained on such a subject that Prout's law . . . is nothing but an illusion", 1860, p. 45.) It is less easy to ascertain how confident Prout and his contemporaries were in the methods by which atomic weights were measured, but it is unlikely that this confidence was very great, in view of the many clear sources of error and the failure of independent measurements generally to produce identical results. On the other hand, chemists of the time must have felt that their methods for determining atomic weights were more likely to be accurate than not, otherwise they would not have used them. For these reasons, we conjecture that $P(a)$ was of the order of 0.6 and that $P(t)$ was around 0.9, and these are the figures we shall work with. It should be stressed that these numbers and those we shall assign to other probabilities are intended chiefly to illustrate how Bayes's Theorem resolves Duhem's problem; nevertheless, we believe them to be sufficiently accurate to throw light on the progress of Prout's hypothesis. As we shall see, the results we obtain are not very sensitive to variations in the assumed prior probabilities.

In order to evaluate the posterior probabilies of t and of a, one must fix the values of the terms $P(e \mid t)$, $P(e \mid a)$ and $P(e)$. These can be expressed, using the Theorem on Total Probability (Chapter 2, section **e**), as follows:

$$P(e) = P(e \mid t)P(t) + P(e \mid {\sim}t)P({\sim}t)$$

$$\begin{aligned} P(e \mid t) &= P(e \,\&\, a \mid t) + P(e \,\&\, {\sim}a \mid t) \\ &= P(e \mid t \,\&\, a)P(a \mid t) + P(e \mid t \,\&\, {\sim}a)P({\sim}a \mid t) \\ &= P(e \mid t \,\&\, a)P(a) + P(e \mid t \,\&\, {\sim}a)P({\sim}a) \end{aligned}$$

Since t & a, in combination, is refuted by e, the term $P(e \mid t \,\&\, a)$ is zero. Hence:

$$P(e \mid t) = P(e \mid t \,\&\, {\sim}a)P({\sim}a).$$

It should be noted that in deriving the last equation but one, we have followed Dorling in assuming that t and a are independent, that is, that $P(a \mid t) = P(a)$ and, hence, $P({\sim}a \mid t) = P({\sim}a)$. This seems to accord with many historical cases and is clearly right in the present case. By parallel reasoning to that employed above, we may derive the results:

$$P(e \mid a) = P(e \mid {\sim}t \,\&\, a)P({\sim}t)$$

$$P(e \mid {\sim}t) = P(e \mid {\sim}t \,\&\, a)P(a) + P(e \mid {\sim}t \,\&\, {\sim}a)P({\sim}a).$$

Provided the following terms are fixed, which we have done in a tentative way, to be justified presently, the posterior probabilities of t and of a can be determined:

$$P(e \mid {\sim}t \,\&\, a) = 0.01$$

$$P(e \mid {\sim}t \,\&\, {\sim}a) = 0.01$$

$$P(e \mid t \,\&\, {\sim}a) = 0.02.$$

The first of these gives the probability of the evidence if Prout's hypothesis is not true but if the method of atomic weight measurement is accurate. Such probabilities were explicitly considered by some nineteenth century chemists, and they typically took a theory of random assignment of atomic weights as the alternative to Prout's hypothesis (e.g., Mallet, 1880); we shall follow this. Suppose it had been established for certain that the atomic weight of chlorine lay between 35 and 36. (The final results we obtain respecting the posterior probabilities of *t* and *a* are, incidentally, not affected by the width of this interval.) The random-allocation theory would assign equal probabilities to the atomic weight of an element lying in any 0.01-wide interval. Hence, on the assumption that *a* is true, but *t* false, the probability that the atomic weight of chlorine lies in the interval 35.825 to 35.835 is 0.01. We have assigned the same value to $P(e \mid {\sim}t \,\&\, {\sim}a)$ on the grounds that if *a* were false because, say, some of the chemicals were impure or the measuring techniques faulty, then, still assuming *t* to be false, one would not expect atomic weights to be biased towards any particular part of the interval between adjacent integers.

We have set the probability $P(e \mid t \,\&\, {\sim}a)$ rather higher, at 0.02. The reason for this is that although some impurities in the chemicals and some degree of inaccuracy in the method of measurement were moderately likely in the early nineteenth century, chemists certainly would not have considered their techniques entirely haphazard. Thus if Prout's hypothesis were true, but the measuring technique imperfect, the measured atomic weights would have been likely to deviate somewhat from integral values; but the greater the deviation, the less likely, on these assumptions, so the probability of an atomic weight lying in any part of the 35–36 interval would not be distributed uniformly over the interval, but would be more concentrated around the whole numbers. Let us proceed with the figures we have assumed for the crucial probabilities.* We thus obtain:

$$\begin{aligned}
P(e \mid {\sim}t) &= 0.01 \times 0.6 + 0.01 \times 0.4 &&= 0.01\\
P(e \mid t) &= 0.02 \times 0.4 &&= 0.008\\
P(e \mid a) &= 0.01 \times 0.1 &&= 0.001\\
P(e) &= 0.008 \times 0.9 + 0.01 \times 0.1 &&= 0.0082
\end{aligned}$$

*As a matter of fact, it is not the particular values taken by the three probability terms that are important, but their *relative* values. Thus we would arrive at the same posterior probabilities for *a* and *t* with the weaker assumptions that $P(e \mid {\sim}t \,\&\, a) = P(e \mid {\sim}t \,\&\, {\sim}a) = \frac{1}{2} P(e \mid t \,\&\, {\sim}a)$.

Finally, Bayes's Theorem enables us to derive the posterior probabilities in which we were interested:

$$P(t \mid e) = 0.878 \quad \text{(Recall that } P(t) = 0.9\text{)}$$

$$P(a \mid e) = 0.073 \quad \text{(Recall that } P(a) = 0.6\text{)}$$

These striking results show that evidence of the kind we have described may have a sharply asymmetric effect on the probabilities of *t* and of *a*. The initial probabilities we assumed seem appropriate for chemists such as Prout and Thomson, and if they are correct, the results deduced from Bayes's Theorem explain why those chemists regarded Prout's hypothesis as being more or less undisturbed when certain atomic-weight measurements diverged from integral values, and why they felt entitled to adjust those measurements to the nearest whole number. Fortunately, these results are relatively insensitive to changes in our assumptions, so their accuracy is not a vital matter as far as our explanation is concerned. For example, if one took the initial probability of Prout's hypothesis *(t)* to be 0.7, instead of 0.9, keeping the other assignments, we find that $P(t \mid e) = 0.65$, while $P(a \mid e) = 0.21$. Thus, as before, after the refutation, Prout's hypothesis is still more likely to be true than false, and the auxiliary assumptions are still much more likely to be false than true. Other substantial variations in the initial probabilities produce similar results, though with so many factors at work, it is difficult to state concisely the conditions upon which these results depend without just pointing to the equations above. Thus Bayes's Theorem provides a model to account for the kind of scientific reasoning that gave rise to the Duhem problem. And the example of Prout's hypothesis, as well as others that Dorling (1979 and 1982) has described, show, in our view, that the Bayesian model is essentially correct. By contrast, non-probabilistic theories seem to lack entirely the resources that could deal with Duhem's problem.

A fact that emerges when slightly different values are assumed for the various probabilities in the Prout's hypothesis example is that one or other of the theories may actually become more probable after the conjunction *t* & *a* has been refuted. For instance, when $P(e \mid t \,\&\, {\sim}a)$ equals 0.05, the other probabilities being assigned the same values as before, the posterior probability of *t* is 0.91, which exceeds its prior probability. This may seem bizarre but, as Dorling (1982) has argued, it is not so odd when one bears in mind that the refuting evidence normally contains a good deal more information than is required merely to disprove *t* & *a* and that this extra information may be confirmatory. In general, such confirmation occurs when $P(e) < P(e \mid t)$, which is easily shown to be equivalent to the condition $P(e \mid t) > P(e \mid {\sim}t)$. In other words, when evidence is easier to explain (in the sense of having a higher probability) if a given hypothesis is true than if it is not, then that theory is confirmed by the evidence.

I GOOD DATA, BAD DATA, AND DATA TOO GOOD TO BE TRUE

Good data. The marginal influence which we have seen an anomalous observation may exert on the probability of a theory is to be contrasted with the dramatic effect that a confirmation can have. For instance, if the measured atomic weight of chlorine had been a whole number, in line with Prout's hypothesis, so that now $P(e \mid t \,\&\, a)$ is one instead of zero, and if the probabilities we assigned were kept, the probability of the hypothesis would have shot up from a prior of 0.9 to 0.998. And, even more dramatically, if the prior probability of t had been 0.7, its posterior probability would have risen to 0.99. The existence of this asymmetry between anomalous and confirming instances was highlighted with particular vigour by Lakatos, who regarded it as being of the greatest significance in science and as one of the characteristic features of a research programme; Lakatos maintained that a scientist involved in such a programme typically "forges ahead with almost complete disregard of 'refutations'", provided he is occasionally rewarded with successful predictions (1970, p. 137): he is "encouraged by Nature's YES, but not discouraged by its NO" (1970, p. 135). As we have indicated, we believe there to be much truth in Lakatos's observations; however, they are merely incorporated without explanation into his methodology, while the Bayesian has a simple and plausible explanatory model.

Bad data. An interesting fact that emerges from the Bayesian analysis is that a successful prediction derived from a combination of two theories, say t and a, does not always redound to the credit of t, even if the prior probability of the evidence is small; indeed, it can even undermine it. We may illustrate this by referring again to the example of Prout's hypothesis.

Suppose the atomic weight of chlorine were 'measured', not in the old-fashioned chemical way, but by concentrating hard on the element in question and picking a number in some random fashion from a given range of numbers. And let us assume that this method assigns a whole-number value to the atomic weight of chlorine. This is just what one would predict on the basis of Prout's hypothesis, if the outlandish measuring technique were reliable. But reliability is obviously most unlikely and it is equally obvious that, as a result, the measured atomic weight of chlorine adds practically nothing to the probability of Prout's hypothesis, notwithstanding its integral value. This intuition is upheld by Bayes's Theorem, as a simple calculation based on the above formulas shows. (As before, let t be Prout's hypothesis and a the assumption that the measuring technique is accurate. Then set $P(e \mid t \,\&\, {\sim}a) = P(e \mid {\sim}t \,\&\, {\sim}a) = P(e \mid {\sim}t \,\&\, a) = 0.01$, for reasons similar to those stated earlier, and let $P(a)$ be very small, say 0.0001, for obvious reasons. It then follows that $P(t)$ and $P(t \mid e)$ are equal to two decimal places.) This example shows that Leibniz was wrong to declare as a general principle that "It is the greatest commendation of an hypothesis (next to truth) if by its help predictions can be made even about phenomena or experiments not tried". Leibniz and Lakatos, who quoted these words with approval (1970, p. 123), seem to have overlooked the fact that if a prediction can be deduced from a hypothesis only with the assistance of highly questionable auxiliary claims, then that hypothesis often accrues very little credit. This explains why the various sensational predictions which Velikovsky drew from his theory of planetary collisions failed to impress most serious scholars, even when some of those predictions were to their amazement fulfilled. For instance, Velikovsky's prediction of the existence of large quantities of petroleum on the planet Venus relied not only on his pet theory that various natural disasters in the past had been caused by collisions between the earth and a comet, but also on a number of unsupported and not very plausible assumptions, such as that the comet in question originally carried hydrogen and carbon, that these had been converted to petroleum by electrical discharges supposedly created in the violent impact with the earth, that the comet had later evolved into the planet Venus, and some others (Velikovsky, 1950, p. 351). (More details of Velikovsky's theory are given in the next section.)

Data too good to be true. Data are sometimes said to be 'too good to be true' when they seem to fit a favoured hypothesis more perfectly than it is reasonable to expect. For instance, suppose all the atomic weights listed in Prout's paper had been whole numbers, exactly. Such a result almost looks as if it was designed to impress, and it is just for this reason that it fails to. We may analyse this response as follows. Let e be the evidence of, say, 20 atomic-weight measurements, each a perfect whole number. No one could have regarded precise atomic weights measured at the time as absolutely reliable. The most natural view would have been that such measurements are subject to experimental error and, hence, that they would give a certain spread of results about the true value. On this assumption, which we shall label a', it is extremely unlikely that numerous independent atomic-weight measurements would all produce whole numbers, even if Prout's hypothesis were true. So $P(e \mid t \,\&\, a')$ is extremely small and, clearly, $P(e \mid {\sim}t \,\&\, a')$ would be no larger. Now a' has many possible alternatives, one of the more plausible (though initially it might not be very plausible) being that the experiments were consciously or unconsciously rigged in favour of Prout's hypothesis. If this were the only significant alternative (and so, in effect, equivalent to ${\sim}a'$), $P(e \mid t \,\&\, {\sim}a')$ would be very high, as would $P(e \mid {\sim}t \,\&\, {\sim}a')$. It follows from the equations on pages 99–100 above that

$$P(e \mid t) \approx P(e \mid t \,\&\, {\sim}a')P({\sim}a') \text{ and}$$
$$P(e \mid {\sim}t) \approx P(e \mid {\sim}t \,\&\, {\sim}a')P({\sim}a')$$

and, hence,

$$P(e) \approx P(e \mid t \,\&\, {\sim}a')P({\sim}a')P(t) + P(e \mid {\sim}t \,\&\, {\sim}a')P({\sim}a')P({\sim}t).$$

Now, presumably the rigging of the results to produce whole numbers, if it took place, would produce whole numbers equally effectively whether t was true or not; in other words,

$$P(e \mid t \,\&\, {\sim}a') = P(e \mid {\sim}t \,\&\, {\sim}a');$$

hence

$$P(e) \approx P(e \mid t \,\&\, {\sim}a')P({\sim}a').$$

Therefore,

$$P(t \mid e) = \frac{P(e \mid t)P(t)}{P(e)} \approx \frac{P(e \mid t \,\&\, {\sim}a')P({\sim}a')P(t)}{P(e \mid t \,\&\, {\sim}a')P({\sim}a')} = P(t)$$

Thus e does not confirm t significantly, even though, in a misleading sense, it fits the theory perfectly. This is why it is said to be too good to be true. A similar calculation shows that the probability of a' is diminished and, on the assumptions that we made, this implies that the probability of the experiments having been fabricated is enhanced. (The above analysis is essentially the same as given in Dorling, 1982).

A famous case of data that were criticized for being too good to be true is that of Mendel's plant-breeding results. Mendel's genetic theory of inheritance allows one to calculate the probabilities with which certain plants would produce specific kinds of offspring. For instance, under certain circumstances, pea plants of a particular strain may be calculated to yield round and wrinkled seeds with probabilities 0.75 and 0.25, respectively. Mendel obtained seed-frequencies that matched the corresponding probabilities in this and in similar cases remarkably well, suggesting (misleadingly Fisher contended) substantial support for the genetic theory. Fisher did not believe that Mendel had deliberately falsified his results to appear in better accord with his theory than they really were. To do so, Fisher claimed, would "contravene the weight of the evidence supplied in detail by . . . [Mendel's] paper as a whole" (1936, p. 132). But Fisher thought it a "possibility among others that Mendel was deceived by some assistant who knew too well what was expected" (1936, p. 132), an explanation he backed up with some (rather meagre) independent evidence.

The argument put forward earlier to show that too-exactly whole-number atomic-weight measurements would not have supported Prout's hypothesis depends on the existence of some sufficiently plausible alternative hypothesis that explains the data better. We believe that, in general, data are too good to be true relative to one hypothesis only if there are such alternatives. This principle accords with intuition; for if the technique for eliciting atomic weights had long been established as precise and accurate, and if careful precautions had been taken against experimenter bias, all the natural alternatives to Prout's hypothesis could be discounted and the data would no longer seem suspiciously good; they would be straightforwardly good. Fisher, however, did not subscribe to the principle, at least, not explicitly; he believed that Mendel's results told against the genetic theory whatever alternative explanations might suggest themselves. Nevertheless, as just indicated, the consideration of such alternatives played a part in his argument. We shall refer again to Fisher's case against Mendel in the next chapter.

j AD HOC HYPOTHESES

As we have seen, an important scientific theory which, in combination with other assumptions, has made a false prediction may nevertheless emerge relatively unscathed, while the auxiliary hypotheses are largely discredited. (We are using such expressions in the normal way to describe how hypotheses are received, regarding them as harmless metaphors for obvious and more or less precise probabilistic notions. Thus, a hypothesis that is unscathed by negative evidence is one whose posterior and prior probabilities are similar. On the other hand, it is difficult to understand what opponents of the Bayesian approach could have in mind when they talk of theories being 'accepted' or 'retained', or 'put forward' or 'saved' or 'vindicated'.) When a set of auxiliary assumptions is discredited in a test, scientists frequently think up new assumptions which assist the main theory to explain the previously anomalous data. Sometimes these new assumptions give the impression that their role is simply to 'patch up' the theory, and in such cases Francis Bacon called them "frivolous distinctions" (1620, Book I, aphorism xxv). More recently they have been tagged 'ad hoc hypotheses', presumably because they would not have been introduced if the need to bring theory and evidence into line had not arisen. However, the term is pejorative, and hypotheses falling into the ad hoc category are very often dismissed as more or less worthless.

But although particular ad hoc theories are fairly easy to evaluate intuitively, there is controversy over what general criteria apply. Indeed, there is not even a universally accepted definition of 'ad hoc' as that term is applied to hypotheses. We shall see that the Bayesian approach clarifies the question. First let us consider a few uncontroversial examples and then deal with some general accounts of ad hocness.

j.1 Some Examples of Ad Hoc Hypotheses

Velikovsky's theory of collective amnesia. Immanuel Velikovsky, in a daring book called *Worlds in Collision* that attracted

a great deal of attention some years ago, put forward the theory that the world has been subject, at various stages in its history, to cosmic disasters produced by near collisions with massive comets. One of these comets, which went on to make a distinguished career as the planet Venus, is supposed to have passed close by the earth during the Israelites' captivity in Egypt and to have caused the various remarkable events of the time, such as the ten plagues and the parting of the Red Sea. One of the theory's predictions, apparently, is that every group of people in the world will have noticed these tremendous goings-on and if they kept records at all, they would have recorded them. However, many communities failed to note in their writings anything out of the ordinary at that time, and Velikovsky, remaining convinced by his main theory, put this exceptional behaviour down to what he called a "collective amnesia". He argued that the cataclysms were so terrifying that whole peoples behaved "as if [they had] obliterated impressions that should be unforgettable". There was a need, Velikovsky said, to "uncover the vestiges" of these events, "a task not unlike that of overcoming amnesia in a single person" (1950, p. 288). Individual amnesia is the issue in the next example.

Dianetics. Dianetics is a theory that purports to analyse the causes of insanity and mental stress, which it sees as the 'misfiling' of information in inappropriate locations in the brain. By refiling these 'engrams', it claims, sanity may be restored, composure enhanced, and, incidentally, the memory vastly improved. Not surprisingly, the therapy is long and expensive, and few people have been through it and borne out the theory's claims. One triumphant success, a young student, was, however, announced by the inventor of Dianetics, L. Ron Hubbard, and in 1950 he exhibited this person to a large audience, claiming that she had a "full and perfect recall of every moment of her life". However, questions from the floor ("What did you have for breakfast on October 3, 1942?"; "What colour is Mr Hubbard's tie?", and the like) soon demonstrated that the hapless girl had a most imperfect memory. Hubbard accounted for this to what remained of the assembly by saying that when the girl first appeared on the stage and was asked to come forward "now", the word "now" had frozen her in "present time" and paralysed her ability to recall the past. (An account of the incident and of the history of Dianetics is given by Miller, 1987.)

An example from psychology. Investigations into distributions of IQ show that different groups of people vary in their average levels of measured intelligence. A number of so-called environmentalists put a low score down primarily to poor social and educational conditions. However, this explanation ran into trouble when it was discovered that a large group of Eskimos, leading a feckless, poor, and drunken existence, scored very highly on IQ tests. The distinguished biologist Peter Medawar (1974), in an effort to deflect the difficulty away from the environmentalist thesis, explained the observation by saying that an "upbringing in an igloo gives just the right degree of cosiness, security and mutual contact to conduce to a good performance in intelligence tests."

In each of these examples, the theory which replaced the refuted one seems rather unsatisfactory. It is not likely that they would have been put forward except in response to a particular empirical anomaly, and this explains the label "ad hoc", which suggests that the theory was advanced for the specific purpose of evading a difficulty. However, some theories of this kind cannot be condemned so readily. For instance, an ad hoc alteration which rescued Newtonian theory from a difficulty led directly to the discovery of a new planet and was generally deemed a great success.

The discovery of the planet Neptune. Newtonians tried unsuccessfully to account for the motion of the planet Uranus, but the difference between theory and observation exceeded the admissible limits of experimental error. Two astronomers, Adams and Leverrier, working independently, put forward a new theory which postulated the existence of a previously unthought-of planet and hence of a new source of gravitational attraction to act on Uranus. This theory was later vindicated by careful telescopic observations and studies of old astronomical maps, which revealed the presence of a planet with the anticipated characteristics. The planet was later called Neptune. (The fascinating story of this episode is told by W. M. Smart, 1947.)

J.2 A Standard Account of Ad Hocness

The salient features of the examples we are considering are that a theory *t*, which we can call the main theory, was combined with an auxiliary hypothesis, *a*, to predict *e*, when in fact

e' occurred, *e'* being incompatible with *e*. And in order to retain the main theory in its desired explanatory role, a new auxiliary, *a'*, was proposed which, with *t*, implies *e'*.

Two criteria of acceptability are often applied by philosophers in such circumstances. The first is that *t* & *a'* should have test implications that are independent of the evidence that refuted *t* & *a*. The second criterion is that some of these test implications should be verified. Lakatos (1970, p. 175) called theories that failed the first, ad hoc$_1$, and those that did not satisfy the second, ad hoc$_2$. Some philosophers maintain that a theory is acceptable only if it is non-ad hoc in both of these senses (for example, Popper, 1963, pp. 244–248), while others emphasise only the first sense (for example, Hempel, 1966, p. 29). Our criticisms of this approach will not need to distinguish between the two points of view.

The term ad hoc to describe hypotheses that do not meet one or other of these conditions seems not to be an old one; its earliest occurrence in English that we know of was in 1936, in a critical review of a book of psychology. The reviewer, W. J. H. Sprott, observed that

> There is a suspicion of '*ad-hoc*-ness' about the 'explanations' [of a certain aspect of childish behaviour]. The whole point is that such an account cannot be satisfactory until we can predict the child's movements from a knowledge of the tensions, vectors and valences which are operative, *independent of our knowledge of how the child actually behaved*. So far we seem reduced to inventing valences, vectors and tensions from a knowledge of the child's behaviour. (Sprott, 1936, p. 249; our emphasis)

But although the term ad hoc is relatively new, the idea goes back at least to Bacon, who criticized as a "frivolous distinction" the type of hypothesis that is "framed to the measure of those particulars only from which it is derived". Bacon argued that a hypothesis ought to be "larger and wider" than the observations that gave rise to it and, moreover, that it should lead to new particulars. According to this criterion, the first three examples above seem to be unsatisfactory scientific developments, while the fourth does not, since the new-planet theory was supported by evidence different from that which led to the original refutation. According to this criterion, the modification which Velikovsky brought to his theory would be acceptable only if it were supported, for example, by contemporary records of amnesia or by evidence of peculiar features in the environment which we have reason to think are conducive to mass forgetfulness. Medawar's and Hubbard's theories are rather vague and seem unsusceptible of any independent test, though one must acknowledge that a closer study of those theories could reveal potential tests.

j.3 A Bayesian Account of Ad Hocness

We shall argue in the next subsection that the above account misses the characteristic of ad hoc hypotheses that determines whether they are well regarded or not by scientists. The quantity which, according to the Bayesian, influences one's evaluation of a scientific development is the posterior probability of the revised theory, and for the theory to be 'acceptable' in the everyday sense of the term, this should be relatively high—at any rate, it ought to exceed 0.5. If a theory is more probable than 0.5, then it is more likely to be true than false, which would seem to be a minimum condition for 'acceptability'. On this view, a' will be judged adversely and pejoratively labelled ad hoc, if $P(a' \mid e' \,\&\, b) \leq 0.5$, where e' is the new evidence that refuted the predecessor of a' and b is any other relevant information. In this account (which agrees with that given by Horwich, 1982, pp. 105–108), there is no need for a' to be supported by evidence independent of e'; all that is wanted is that it be credible. Scientists are also interested in whether t in the presence of the newly-thought-up a' provides a competent explanation of the previously anomalous e'. It would do so only if $t \,\&\, a'$ was a sufficiently credible theory; since $P(t \,\&\, a' \mid e' \,\&\, b) \leq P(a' \mid e' \,\&\, b)$, this would be the case only if a' were not ad hoc.

The Bayesian account explains the low esteem which ad hoc hypotheses frequently command in the scientific community. It also explains why people often respond with immediate incredulity, indeed derision, to an ad hoc hypothesis. Is it likely that their amusement comes from perceiving that the hypothesis leads to no new predictions? We do not believe so. Finally, the Bayesian account explains why the hypotheses are termed ad hoc. For since an ad hoc hypothesis was originally improbable, it would not have been seriously entertained if e', the evidence that undermined an earlier hypothesis, had not been discovered, and the need to explain the new anomalous result had not arisen.

j.4 Why the Standard Account Must Be Wrong

The standard account characterises ad hoc hypotheses as being unsupported or unsupportable by evidence independent of that which led to their being proposed. It is implicit in that account that only hypotheses that do enjoy such support are acceptable. We shall argue that this is wrong, both in the light of counter-examples and by means of a more general argument. The standard account of ad hocness is, no doubt, inspired in part by the desire to avoid attributing inevitably subjective probabilities to theories. If so, the aim backfires, for, as we shall show, the non-Bayesian account has its own subjective aspect, one which, in our view, is very inappropriate.

Consider first a couple of counter-examples to the standard account. Suppose one were examining the hypothesis that an urn contains only red counters. An experiment is conducted in which counters are removed at random and then replaced, and this trial is repeated, say, 10,000 times. Let the result of this trial be that 4950 of the selected counters were red and the rest white. The initial hypothesis, and the various necessary auxiliary assumptions, are together refuted, and a natural revision would be that the urn contains red and white counters in approximately equal numbers. This seems a perfectly legitimate procedure, and the revised hypothesis appears well justified by the evidence, yet there is no independent evidence for it: its support comes solely from the evidence which discredited its predecessor (Howson, 1984).

Theorising about the contents of an urn is only a humble form of enquiry, but there is no reason to think that these conclusions do not also hold in the higher sciences. Indeed, the following is a case where they do hold: the assumption is made that two characteristics of a plant are inherited in accordance with Mendel's principles and that each is controlled by a specific gene, the two genes acting independently and being located on different chromosomes. The results of plant-breeding experiments show that a surprising number of plants carry both characteristics, and the original assumption that the genes act independently is revised in favour of a theory that they are linked on the same chromosome. Again, the revised theory would be strongly confirmed and established as acceptable merely on the evidence that stimulated its formulation and without the necessity of further, independent, evidence. (An example of this sort is worked out by Fisher in his *Statistical Methods for Research Workers*, ch. IX.)

We turn now to a more general objection to the idea that hypotheses are acceptable only if corroborated by independent evidence. Imagine a scientist who performs an experiment and observes e', which because it implies the falsity of the prediction e made by t & a, refutes that combination of theories. Suppose a new theory, t & a', is advanced, which is ad hoc in one or other of the two senses that there is either no fresh evidence for a' or no possibility of such evidence. The theory therefore is unacceptable, according to the view we are considering. But this cannot be so. For consider that only part of the observational evidence, namely $\sim e$, is required for the refutation. Now suppose another scientist first contrived an experiment with only two possible outcomes: either e or $\sim e$. Having obtained the latter, he revises his theory to t & a', performs the orthodox experiment, and observes e'. In cases where e' is not implied or made highly probable by $\sim e$, according to the view we are discussing, this new theory would be perfectly acceptable since it is supported by evidence independent of that which refuted its predecessor. But the two experimenters are in precisely the same position as regards the available evidence, yet for the one the theory is unacceptable, for the other it is not! This is a most alarming consequence for any methodology, for it fails spectacularly to reflect scientific reasoning and flies in the face of common sense.

It is curious, too, that a methodology designed to provide purely objective criteria should lead to the conclusion that the epistemic value of a theory is so closely connected with the state of mind of its inventor.

4.5 The Notion of Independent Evidence

As we have explained, the non-inductivist, non-Bayesian account of ad hocness asserts that a theory consisting of the combination t & a is only replaced by t & a' in an acceptable scientific fashion when a' is successfully tested by evidence independent of that which refuted the first theory. This thesis is often associated with another, rather similar, view, namely, that no theory is acceptable unless it is supported by evidence independent of that which prompted its initial proposal, whether this also refuted a predecessor or not. We have shown that neither of these views is either reasonable or compatible with scientific practice, and, moreover, that they fail to deliver the objective standards of theory-appraisal to which they aspire. (Howson, 1984, addresses a number of other objections.) One problem with the non-Bayesian criterion of ad hocness, which we have not needed to exploit in our criticism of it, is that the notion of 'independent' evidence is left vague and intuitive. Moreover, there seems to be no way of interpreting the notion in a purely objective fashion.

It seems not to be the probabilistic sense of independence that is intended. For suppose $P(e_2 \mid e_1) < P(e_2)$. This means that e_2 is not probabilistically independent of e_1. Nevertheless, if $P(e_2 \mid e_1)$ were sufficiently small, e_2 would (it is generally acknowledged) support an appropriate theory, even if e_1 were already known and had been counted in support of the theory. Hence, the notion of independence that is often employed in this context cannot be the probabilistic notion. Another possibility would be that e_1 is independent of e_2 just in case neither entails the other. But this would mean that if the two bits of evidence were trivially distinct in, say, relating to different times, or slightly different places, then they would be independent. And this would mean that practically no theory would be ad hoc. For instance, Medawar's peculiar theory about the cosiness of the Eskimo's way of life was propounded in response to some surprising IQ measurements that had been reported. Presumably, one could infer from the theory that IQ tests applied the following week to the same group of Eskimos would produce similar results. But although this prediction is logically independent of the earlier reported results, it would not significantly improve the standing of Medawar's theory.

What seems to be wanted of evidence in the standard account for it to save a hypothesis from ad hocness is that it should be supported by evidence that is *different* from that which led to its predecessor's downfall. Indeed, the ideas of dependence and independence seem closely related to those of similarity and diversity, so we shall continue the discussion by considering these notions.

Evidence that is varied is often regarded as offering better support to a hypothesis than an equally extensive volume of homogeneous evidence. As Hempel put it, "the confirmation of a hypothesis depends not only on the quantity of the favorable evidence available, but also on its variety: the greater the variety, the stronger the resulting support" (1966, p. 34). According to the Bayesian, if two sets of data are entailed by a hypothesis (or have similar probabilities relative to it) and one of

Variety
of
ev

them confirms that hypothesis more than the other, this must be due to a corresponding difference between the data in their probabilities. In other words, the variety of evidence is a matter of its probability. We shall explain.

Consider first some examples: the report of the rate at which a stone falls to earth from a given height on a Tuesday is similar to that relating to the stone's fall on a Thursday, say; it is very different, however, to a report of the trajectory of a planet or one of the manner in which a given fluid rises in a capillary tube; each of these reports, however, confirms Newton's theory, though to varying degrees. The similar instances in the above list have the characteristic that when one of them is known, any other would thereby be anticipated with high probability. This recalls Francis Bacon's characterisation of similarity in the context of inductive evidence. He spoke of observations "with a promiscuous resemblance one to another, insomuch that if you know one you know all" and was probably the first to point out that it would be superfluous to cite more than a small representative sample of such observations in evidence (*see* Urbach, 1987, pp. 160–164). The idea of similarity between items of evidence is expressed naturally in probabilistic terms by saying that e_1 and e_2 are similar if $P(e_2 \mid e_1)$ is higher than $P(e_2)$ and one might add that the more the first probability exceeds the second, the greater the similarity. This means that e_2 would provide less support if e_1 had already been cited as evidence than if it was cited by itself.

On the other hand, knowing that one of a pair of dissimilar instances has occurred gives little or no guidance as to whether the other will occur. For example, unless Newton's, or some comparable theory, had already been firmly established, a knowledge of the rate of fall of a given object on some specific occasion would not significantly affect one's confidence that the planet Venus, say, would appear in a particular position in the sky on a designated day. Different pieces of evidence may also have a mutually discrediting effect. An example of this might be the observations of the same constant acceleration of heavy bodies dropped at sea level and the unequal rates of fall of objects released on different mountaintops. Both observations would confirm Newton's laws, but in circumstances where those laws are not already well established, the first set of observations might suggest that all objects falling freely (whether on top of a mountain or not) do so with the same acceleration. In

other words, with different instances, say e_3 and e_1, $P(e_3 \mid e_1)$ is either close to or less than $P(e_3)$. Of course, e_3 merely being different from e_1 in this sense does not imply that it supports any hypothesis significantly; whether it does or not depends on its probability. The notion of similarity, as we have characterised it, is reflexive, as it should be; that is, if e_2 is (dis)similar to e_1, then e_1 is (dis)similar to e_2 (this follows directly from Bayes's Theorem).

Our characterisation of similarity and diversity in data closely resembles the one we gave earlier in relation to experiments (*see* section **d**). They are not identical, however, since there is nothing in the earlier definitions to preclude the results of one particular experiment counting as a heterogeneous set of data. For an experiment is defined simply in terms of a set of instructions. If such instructions said, for example, 'first throw a die, then if a six appears, measure the speed of light, if a five, ascertain the position of the planet Mars, if a four, . . .'one might obtain varied data, all from the same experiment. It is natural to respond by saying that this complex experiment really comprises a variety of different ones and demand a definition of 'experiment' in terms of the homogeneity of its outcomes rather than in terms of a set of practical directions. Such a definition is no doubt possible, but we feel it is bound to be somewhat arbitrary and unrewarding; it would, for instance, have to stipulate exactly how homogeneous the outcomes of some set of operations should be in order to qualify as a unitary experiment.

■ k INFINITELY MANY THEORIES COMPATIBLE WITH THE DATA

k.1 The Problem

Galileo carried out many experiments on freely falling bodies and on bodies rolling down inclined planes in which he examined how long they took to descend various distances. These experiments led him to formulate the well-known law to the effect that $s = ut + \frac{1}{2}gt^2$, where s is the distance fallen by a freely falling body, u is its initial downward velocity, g is a constant, and t is the time taken by the fall. Jeffreys (1961, p. 3) pointed out that Galileo might also have advanced the following as his law:

$$s = ut + \frac{1}{2}gt^2 + f(t)\,(t - t_1)(t - t_2)\ldots(t - t_n),$$

where $t_1, t_2, \ldots,$ and t_n represent the times at which he carried out his experiments, and where f is any function of t at all. Thus Jeffreys's modification stands for an infinite number of alternatives to Galileo's theory. Although all these theories contradict one another and make different predictions about future experiments, the interesting feature of Jeffreys's unorthodox laws of free fall is that they all imply those data which Galileo had from his experiments.

This is hard to reconcile with those non-inductivist, non-probabilistic theories of scientific method which hold that the scientific value of a theory is determined just by the evidential support it has, where that support is simply a function of $P(e \mid h)$ and, in some versions, of $P(e)$. These philosophical approaches would have to regard the standard law of free fall and those peculiar alternatives described by Jeffreys as equally good scientific theories, relative to the evidence available to Galileo, although this is a judgment with which no scientist would agree. The same point emerges from a well-known example due to Nelson Goodman (1954). He noted that the evidence of very many and varied green emeralds would normally suggest that all emeralds are green. But he pointed out that that evidence bears the same relation to "All emeralds are green" as it does to a type of hypothesis he formulated as "All emeralds are grue". According to Goodman's definition, something is *grue* if it is either green and observed before time *t*, or blue and observed at or later than *t*. If *t* denotes some time after the emeralds described in the evidence were observed, then both the green- and the grue-hypotheses imply that the observed emeralds should be green. However, the hypotheses are incompatible, differing in their predictions about the colour of emeralds looked at after the critical time. As with Jeffreys's variants of Galileo's theory, the grue-hypothesis represents an infinite number of alternatives to the more natural hypothesis, for *t* can assume any value, provided it is later than now.

Our examples illustrate a general problem for methodology: that a theory which explains (in the sense of implying or associating a certain probability with) some data is merely one out of an infinite set of rival theories, each of which does the same. The existence of this infinite set of possible explanations, it will be remembered, spelled ruin for any attempt at a positive solution to the problem of induction (*see* Chapter 1). The problem with which we are concerned here arises because, in practice, scientists discriminate between possible explanations and typically pick out just one, or at any rate relatively few, as meriting serious attention. An account of scientific method ought to explain how and why they do this.

k.2 The Bayesian Approach to the Problem

This has not proved easy. For the Bayesian, the nature of the problem, at least, is straightforward. Moreover, Bayesian theory does not imply that every hypothesis similarly related to the data is of equal merit. Suppose one were comparing two theories in the light of the same evidence. Their relative posterior probabilities are given by

$$\frac{P(h_1 \mid e)}{P(h_2 \mid e)} = \frac{P(e \mid h_1)P(h_1)}{P(e \mid h_2)P(h_2)}.$$

If both theories imply the evidence, then $P(e \mid h_1) = P(e \mid h_2) = 1$. And if, in addition, $P(h_1 \mid e)$ exceeds $P(h_2 \mid e)$, then it follows that $P(h_1)$ is larger than $P(h_2)$. More generally, if two theories which explain the data equally well nevertheless have different posterior probabilities, then they must have had different priors too. So theories such as the contrived alternatives to Galileo's law and Goodman's grue-variants must, for some reason, have lower prior probabilities. Indeed, this is clearly reflected in most people finding such hypotheses quite unbelievable. The problem then is to discover the criteria and rationales by which theories assume particular prior probabilities.

Sometimes there is a clear reason why a theory is judged improbable. For instance, suppose the theory concerned a succession of events in the development of a society; it might perhaps assert that the elasticity of demand for herring is a constant or that all future British prime ministers' surnames will start with the letter *T*. These theories, which of course could be true, are however monstrously improbable. And the reason for this is that the events they describe are influenced by numerous independent processes whose separate outcomes are improbable. The probability that all these processes will turn out to favour the hypotheses in question is therefore the product of many small probabilities, and so itself is very small indeed (Urbach, 1987b). The question, of course, remains of how the probabilities of the causal factors are estimated. This could be answered by reference to other probabilities, in which case the question is just pushed one stage back, or else by some different process that does not depend on probabilistic reasoning. For instance, the simplicity of a hypothesis has been thought to have an influence on its initial probability. This and other possible determinants of initial probabilities are discussed in Chapter 11.

It is worth mentioning here that the equation given above, relating the posterior probabilities of two theories with their prior probabilities, explains an important feature of inductive reasoning. The scientist often prefers a theory which explains the data imperfectly, in that $P(e \mid h_1) < 1$, to an alternative, h_2, which predicts them with complete accuracy. Thus, even Galileo's data were not in precise conformity with his theory; nevertheless he did not consider any more complicated function of u and t to be a better theory of free fall than his own, even though it could have embraced the evidence he possessed more perfectly. According to the above equation, this is because the better explanatory power of the rival hypotheses was offset by their inferior prior probabilities (*see* Jeffreys, 1961, p. 4).

■ | CONCLUSION

Charles Darwin (1868, vol. 1, p. 8) said that "In scientific investigations it is permitted to invent any hypothesis, and if it explains various large and independent classes of facts it rises to the rank of a well-grounded theory". This is, perhaps, an exaggeration, for not any hypothesis would do; the hypothesis must not be refuted, or substantially disconfirmed, nor should it be intrinsically too implausible. With these provisos, Bayesianism, we suggest, is just such a well-grounded hypothesis as Darwin referred to. As we showed in Chapter 3, it arises from natural and intuitively reasonable attitudes to risk and uncertainty. It is neither refuted nor undermined by any of the phenomena of scientific reasoning. On the contrary, as we have seen, it explains a wide variety of them. So far we have concentrated chiefly on deterministic theories. We shall see in the next and following chapters that the Bayesian approach is no less successful when dealing with statistical reasoning.

■ PART III

Classical Inference in Statistics

We showed in Chapter 4 how numerous aspects of scientific reasoning can be illuminated by reference to Bayes's Theorem. We confined the discussion there to deterministic theories. As already explained, however, scientific theories are often not deterministic, but are statistical or probabilistic in character. The evaluation of such hypotheses brings no special problems of principle to a Bayesian analysis, the difference between the cases of deterministic and statistical hypotheses being reflected in the term $P(e \mid h)$ which appears in Bayes's Theorem. In the former case, when h entails e, this term equals 1. When h is statistical, $P(e \mid h)$ takes a value equal to the statistical probability which h confers on e, this being an application of the so-called Principal Principle, which is discussed in Chapter 9. Inductive reasoning about deterministic and probabilistic hypotheses is then explained in a uniform fashion in the Bayesian approach, the former merely constituting a special case of the latter.

No such uniform treatment is afforded, however, by the leading non-Bayesian approaches. As a result, a distinct branch of non-Bayesian, or "classical", statistical methodology has grown up since the 1920s. We shall follow the pattern thus established by dealing separately with statistical and deterministic hypotheses. This plan is justified since classical sta-