SCIENTIFIC REASONING
THE BAYESIAN APPROACH

Colin Howson and Peter Urbach

THIRD EDITION

. . . if this [probability] calculus be condemned, then the whole of the sciences must also be condemned.
—Henri Poincaré

Our assent ought to be regulated by the grounds of probability.
—John Locke

OPEN COURT
Chicago and La Salle, Illinois

To order books from Open Court, call toll-free 1-800-815-2280, or visit our website at www.opencourtbooks.com.

Open Court Publishing Company is a division of Carus Publishing Company.

First printing 2006

Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Howson, Colin.
Scientific reasoning : the Bayesian approach / Colin Howson and Peter Urbach.—3rd ed.
p. cm.
Includes bibliographical references (p.) and index.
ISBN-13: 978-0-8126-9578-6 (trade pbk. : alk. paper)
ISBN-10: 0-8126-9578-X (trade pbk. : alk. paper)
1. Science--Philosophy. 2. Reasoning. 3. Bayesian statistical decision theory.
I. Urbach, Peter. II. Title.
Q175.H87 2005
501 dc22

2005024868

Contents

Preface to the Third Edition

How should hypotheses be evaluated, what is the role of evidence in that process, what are the most informative experiments to perform? Questions such as these are ancient ones. They have been answered in various ways, often exciting lively controversy, not surprisingly in view of the important practical implications that different answers carry. Our approach to these questions, which we set out in this book, is the Bayesian one, based on the idea that valid inductive reasoning is reasoning according to the formal principles of probability.

The Bayesian theory derives from the *Memoir* of the mathematician and divine, Thomas Bayes, which was published posthumously by his friend Richard Price in 1763. The principles set out by Bayes had considerable influence in scientific and philosophical circles, though worries about the status of the prior probabilities of scientific theories meant that the whole approach continued to be dogged by debate. And by the 1920s, an alternative approach, often called 'Classical', achieved dominance, due to powerful advocacy by R. A. Fisher and many other distinguished statisticians, and by Karl Popper and similarly distinguished philosophers. Most of the twentieth century was dominated by the classical approach, and in that period Bayesianism was scarcely taught in universities, except to disparage it, and Bayesians were widely dismissed as thoroughly misguided.

But in recent years, there has been a sea-change, a paradigm shift. A search of the Web of Science database shows, during the 1980s, a regular trickle of around 200 articles published annually with the word or prefix 'Bayes' in their titles. Suddenly, in 1991, this number shot up to 600 and by 1994 exceeded 800; by 2000 it had reached almost 1,400. (Williamson and Corfield, 2001, p. 3). This book was one of the first to present a comprehensive,

philosophical case for the Bayesian approach to scientific reasoning and to show its superiority over the classical. Its first and second editions were published in 1989 and 1993, and from the figures quoted it is clear that the book anticipated a powerful and sophisticated resurgence of the once-dominant Bayesian approach.

This new edition amends, updates, re-organizes, and seeks to make the subject more accessible. The text is intended to be self-contained, calling, in the main, on only elementary mathematical and statistical ideas. Nevertheless, some parts are more complex, and some more essential to the overall argument than others. Accordingly, we would suggest that readers who are not already familiar with mathematical probability but who wish to gain an initial understanding of the Bayesian approach, and to appreciate its power, adopt the following plan of attack. First, read Chapter 1, which sets the scene, as it were, with a brief historical overview of various approaches to scientific inference. Then, look at section a of Chapter 2, which gives the simple principles or axioms of the probability calculus, and section b, where there are some of the probability theorems that will be found useful in the scientific context: the central theorem here is Bayes's theorem in its various forms. We then suggest that the reader look at the first few sections of Chapter 4, where Bayes's theorem is applied to some simple reasoning patterns that are found particularly when deterministic theorems are handled; this chapter also compares the Bayesian approach with some others, such as Popper's well-known falsificationist methodology. Chapters 5 to 7 deal with non-deterministic, that is, statistical hypotheses, giving a critical exposition of the classical, or frequentist, methods that constitute the leading alternative to the Bayesian approach; the main classical ideas can be gleaned from sections a to d and f and g of Chapter 5. The final part of the mini-course we are suggesting is to examine Chapter 9, where some of the more widespread criticisms that have been levelled against the Bayesian approach are discussed (and rejected).

There are some marked differences between this third edition and the preceding ones. For example, some of the objections to the Bayesian theory we considered in the second edition have not stood the test of time. There have also been changes of mind: one

of the most prominent examples is the fact that now we accept the strength of de Finetti's well-known arguments against countable additivity, and have accordingly dropped it as a generally valid principle. Other changes have been largely dictated by the desire to make this edition more compact and thereby more accessible. We hope that this indeed turns out to be the case.

CHAPTER 1
Introduction

1.a | The Problem of Induction

Scientific hypotheses have a general character relative to the empirical observations they are supposed to explain, carrying implications about phenomena and events that could not possibly figure in any actual evidence. For instance, Mendel's genetic theory concerns all inherited traits in every kind of flora and fauna, including those that are now extinct and those that are yet to evolve. There is therefore a logical gap between the information derived from empirical observation and the content of typical scientific theories. How then can such information give us reasonable confidence in those theories? This is the traditional Problem of Induction.

One answer that has been suggested claims that our stock of information is not in fact restricted to the empirical. A number of philosophers have taken the view that there are certain principles which are sufficiently rich to bridge the logical gap between observations and theories, whose truth we are able to cognize *a priori*. Kant, for example, held the proposition 'every event has a cause' to be such a principle and he devoted much space and difficult argumentation to proving that it was indeed *a priori*. But whether or not the argument is valid is beside the point, because the principle would not solve the problem of induction anyway. That problem is not essentially concerned with causality; and where specifically causal theories are at issue, the question is not whether every event has a cause, but what the particular cause or causes of a particular observed effect are. Kant (1783, p. 9) tells us that his "dogmatic slumber" was disturbed by Hume's brilliant analysis of the problem of induction, yet he seems not to have fully woken up to its significance.

Another bridging principle that has been proposed is the so-called Principle of the Uniformity of Nature, which Hume (1777, Section 32) summed up in the phrase "the future will resemble the past". Some philosophers have held that when scientists defend their theories, they are tacitly relying on this principle. But it too cannot help with induction. The problem is that the principle is empty, since it does not say in what particular respects the future and the past are similar. And if it is to connect particular observations with a particular theory, it needs a more specific formulation. For example, in order to act as a bridge between observations that certain metals expanded on certain occasions when they were heated and the general proposition that those metals will expand when they are heated in future, the principle needs to be framed in terms of those particular properties. And to infer that *all* metals expand when they are heated would require a more elaborate formulation still. But, as Hume observed, such versions of the Uniformity of Nature Principle are themselves general empirical propositions, whose own claims are no less problematic than the theories they are designed to guarantee.

1.b Popper on the Problem of Induction

It seems, then—and this is no longer controversial—that there is no solution to the problem of induction that could demonstrate with logical certainty the truth of general scientific theories. Some, like Paul Feyerabend, have concluded from the fact that no theory is conclusively proved that all theories are therefore equally unproved, and epistemically on a par, and that the trust we commonly repose in science is completely irrational.

But Karl Popper, amongst others, was concerned to resist such a sweeping scepticism, whose consequences, if accepted, would be alarming. In his attempt to defend the rationality of science and to solve the problem of induction, he drew upon two familiar logical facts. First, that while scientific theories cannot be decisively proved from observational evidence, observations may sometimes refute them. Popper's strong emphasis of this possibility explains why his philosophy is known as Falsificationism. The second logical fact that Popper drew on is that deductive consequences of a

theory can sometimes be verified through observation; when this occurs, Popper said that the theory was "corroborated". This terminology suggests that corroboration confers some epistemic merit on the theory, though it is not clear what implications Popper thought that had for its rational appraisal. The predominant opinion now is that no such implications exist, for when a particular theory is corroborated (in Popper's sense) by evidence, so are infinitely many other theories, all rivals to it and to each other. Only one of these can be true. But which?

Suppose, for example, you were interested in the general law governing the coloration of swans. If the number of swans that will ever exist is n and the number of colours is m, then there are m^n colour combinations. This then represents a lower limit for the number of theories concerning the colours of swans. If we take account of the further possibilities that these birds alter their hues from time to time, and from place to place, and that some of them are multicoloured, then it is apparent that the number of possible theories is immense, indeed, infinite. The simple theory 'all swans are white' that Popper often used as an illustration is corroborated, as he said, by the observation on particular occasions of white swans; but so are infinitely many of the other swan hypotheses. The question of how to support a rational preference amongst these hypotheses then remains. And it is evident that Popper's ideas do nothing to solve the problem of induction.[1]

1.c Scientific Method in Practice

Popper's idea that unrefuted but corroborated hypotheses enjoy some special epistemic virtue, led him to recommend that scientists should seek out and give preference to such hypotheses. There was also a descriptive aspect to this recommendation, for Popper assumed that mainstream science is conducted more or less as he believed it ought to be.

[1] Other examples illustrating the same point are given in Chapter 4, Section i. See also, e.g., Lakatos 1974, Salmon 1981, and Howson 2000 for decisive criticisms of Popper's views on induction.

Popper's descriptive account does reflect two key features of scientific reasoning. First, it sometimes happens in scientific work that a theory is refuted by experimental findings, and when this happens, the scientist usually recognizes that the theory is therefore false and abandons it, perhaps re-adopting it in some revised form. And secondly, when investigating a deterministic theory, scientists frequently focus attention on certain of its logical consequences and then check these empirically, by means of specially designed experiments; and if such consequences turn out to be true, the practice is to conclude that the theory has been confirmed and made more credible.

But Popper's methodology has very little further explanatory power. This is for two reasons. First, it has no means of discriminating between a particular theory that has been confirmed by a successful empirical prediction and the infinity of other, conflicting theories that make the same prediction. In practice, scientists do rank such hypotheses according to their value, or credibility, or eligibility for serious consideration—the scientific enterprise would be impossible otherwise. Secondly, most scientific evidence does not bear the logical relationship to theories that Popper envisaged, for, more usually, such evidence is neither implied by the theories they confirm, nor precluded by those they disconfirm.

So, for example, many deterministic theories that appear in science, especially the more significant ones, often have no directly checkable deductive consequences, and the predictions by which they are tested and confirmed are necessarily drawn only with the assistance of auxiliary theories. Newton's laws, for instance, concern the forces that operate between objects in general and with the mechanical effects of those forces. Observable predictions about particular objects, such as the Earth's planets, can be derived only when the laws are combined with hypotheses about the positions and masses of the planets, the mass-distribution of space, and so on. But although they are not immediate logical consequences of Newton's theory, planetary observations are standardly taken to confirm (and sometimes disconfirm) it.[2]

[2] This objection to Popper's account was pressed with particular effect by Lakatos, as we discuss in Chapter 4.

Then there are scientific theories that are probabilistic, and for that reason have no logical consequences of a verifiable character. For example, Mendel's theory of inheritance states the probabilities with which certain gene combinations occur during reproduction, but does not categorically rule out nor definitely predict any particular genetic configuration. Nevertheless, Mendel obtained impressive confirmation from the results of his plant trials, results which his theory did not entail but stated to be relatively probable.

Finally, even deterministic theories may be confirmed or disconfirmed by evidence that is only assigned some probability. This may arise when a theory's quantitative consequences need to be checked with imperfect measuring instruments, subject to experimental error. For example, the position of a planet at a certain time will be checked using a telescope whose readings are acknowledged in experimental work not to be completely reliable, on account of various unpredictable atmospheric conditions affecting the path of light to the telescope, as well as other uncontrollable factors, some connected with the experimenter and some with physical vagaries. For this reason, quantitative measurements are often reported in the form of a range of values, such as $a \pm b$, where a is the recorded reading and $a - b$ and $a + b$ are the bounds of an interval in which it is judged the true value very probably lies. This judgment is usually based on a theory giving the probabilities that the instrument reading diverges by different amounts from the true value of the measured quantity. Thus, for many deterministic theories, what may appear to be the checking of logical consequences actually involves the examination of experimental effects which are predicted only with a certain probability.

Popper tried to extend his falsificationist ideas to the statistical realm, but insuperable difficulties stand in the way of any such attempt, as we show in Chapter 5. In that chapter, we shall review the ideas of R.A. Fisher, the eminent statistician, who was also inspired by the idea that evidence may have a decisive negative impact on a statistical hypothesis, akin to its falsification. He called a statistical hypothesis under test the 'null hypothesis' and expressed the view that

> the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be

> said to exist only in order to give the facts a chance of disproving the null hypothesis. (1947, p. 16)

Fisher's theory of so-called 'significance tests', which prescribes how statistical hypotheses should be tested, has drawn considerable criticism, and other theories have been advanced in opposition to it. Notable amongst these is the modified theory of significance testing due to Jerzy Neyman and Egon Pearson. Though they rejected much of Fisher's methodology, their theory owed a good deal to his work, particularly to his technical results. Above all, they retained the idea of bivalent statistical tests in which evidence determines one of only two possible conclusions, that is, the acceptance or rejection of a hypothesis.

1.d | Probabilistic Induction: The Bayesian Approach

One of the driving forces behind the development of the above-mentioned methodologies was the desire to vanquish, and provide an alternative to, the idea that the theories of science can be and ought to be appraised in terms of their 'probabilities'. This 'probabilistic induction' is in fact a well-established position in science and philosophy. It has long been appreciated that scientific theories extend beyond any experimental data and hence cannot be verified (that is, logically entailed) by them; yet while it is agreed that absolute certainty is therefore unavailable, many scientists believe that the explanations they think up can secure for themselves an epistemic status somewhere between the two extremes of certainly right and certainly wrong and that that status depends on the quality of the evidence and can be altered by new evidence.

This spectrum of degrees of certainty has traditionally been characterised as a spectrum of probabilities. For example, Henri Poincaré, the noted mathematician and physicist, asked himself what right he had as a scientist to enunciate a theory such as Newton's laws, when it may simply be chance that they are in agreement with all the available evidence. How can we know that the laws will not break down entirely the next time they are tested? "To this objection the only answer we can give is: It is very

improbable" (1905, p. 186). Poincaré believed (pp. 183–84) that "the physicist is often in the same position as the gambler who reckons up his chances. Every time that he reasons by induction, he more or less consciously requires the calculus of probabilities." And summing up his approach, Poincaré remarks (p. 186): "if this calculus be condemned, then the whole of the sciences must also be condemned."

Similarly, the philosopher and economist W.S. Jevons:

> Our inferences . . . always retain more or less of a hypothetical character, and are so far open to doubt. Only in proportion as our induction approximates to the character of perfect induction does it approximate to certainty. The amount of uncertainty corresponds to the probability that other objects than those examined may exist and falsify our inferences; the amount of probability corresponds to the amount of information yielded by our examination; and the theory of probability will be needed to prevent us from over-estimating or under-estimating the knowledge we possess. (1874, Volume 1, p. 263)

Many scientists have voiced the same idea, namely, that theories have to be judged in terms of their probabilities in the light of the evidence. Here are two quotations from Einstein which explicitly manifest a probabilistic view:

> I knew that the constancy of the velocity of light was something quite independent of the relativity postulate and I weighted which was the more probable. (From a letter quoted in Stachel 1998)

and

> Herr Kaufmann has determined the relation [between electric and magnetic deflection] of β-rays with admirable care . . . Using an independent method, Herr Planck obtained results which fully agree [with the computations of] Kaufmann . . . it is further to be noted that the theories of Abraham and Bucherer yield curves which fit the observed curve considerably better than the curve obtained from relativity theory. However, in my opinion, these theories should be ascribed a rather small probability because their basic postulates concerning the mass of the moving electron are not made plausible by theoretical systems which encompass wider complexes of phenomena. (Quoted in Pais 1982, p. 159)

Einstein is here using a basic probabilistic idea, that a very high likelihood (given by the closeness of fit to the data) can combine with a small enough prior probability to give an overall small posterior probability. In fact, as Jon Dorling (1979, p. 180) observed, it is rare to find any leading scientist writing in, say, the last three hundred years who did not employ notions of probability when advocating his own ideas or reviewing those of others.

Philosophers from James Bernoulli in the seventeenth century to Rudolf Carnap, Harold Jeffreys, Bruno de Finetti, Frank Ramsey and E. T. Jaynes in the twentieth century have attempted to explicate these intuitive notions of inductive probability. There have been two main strands in this programme. The first regards the probabilities of theories as objective, in the sense of being determined by logic alone, independent of our subjective attitudes towards them. The hope was that a way could be found to ascertain by logical analysis alone the probability that any given theory is true, and so allow comparative evaluations of competing theories to be placed on an objective footing. This would largely solve the problem of induction and establish an objective and rational basis for science. But, in fact, it is now generally acknowledged that no one has succeeded in this approach, and that objections that have been made against it have proved crippling and unanswerable.

The other strand in the programme to explicate inductive probability treats the probability of a theory as a property of our attitude towards it; such probabilities are then interpreted, roughly speaking, as measuring degrees of belief. This is called the *subjectivist* or *personalist* interpretation. The scientific methodology based on this idea is usually referred to as the methodology of *Bayesianism*, because of the prominent role it reserves for a famous result of the probability calculus known as Bayes's theorem.

Bayesianism has experienced a strong revival in recent years, due in part to its intrinsic plausibility and in part to the weaknesses which have gradually been exposed in the standard methodologies. It is fair to claim that we are in the midst of a revolution, in which Bayesianism is becoming the new paradigm for inductive inference. In the chapters to come, we shall present a detailed account and defence of the Bayesian methodology and will show how it illuminates the various aspects of scientific reasoning.

1.e | The Objectivity Ideal

The sharpest and most persistent objection to the Bayesian approach is directed at one of its defining features, namely that it allows certain subjective factors a role in the appraisal of scientific theories. Our reply to this objection will be that the element of subjectivity is, first of all, minimal and, secondly, exactly right. Such a response contradicts an influential school of thought that denies that any subjectivity at all should be admitted in theory-appraisal; such appraisal, according to that school, should be completely objective. Lakatos (1978, Volume 1, p. 1) expressed this objectivist ideal in uncompromising style, thus:

> The *cognitive* value of a theory has nothing to do with its *psychological* influence on people's minds. Belief, commitment, understanding are states of the human mind. But the objective, scientific value of a theory is independent of the human mind which creates it or understands it, its scientific value depends only on what *objective* support these conjectures have in *facts*.[3]

It was the ambition of Popper, Lakatos, Fisher, Neyman and Pearson, and others of their schools to develop this idea of a criterion of scientific merit, which is both objective and compelling, and yet non-probabilistic. And it is fair to say that their methodologies, especially those connected with significance testing and estimation, which comprise the bulk of so-called Classical methods of statistical inference, have achieved preeminence in the field and become standards of correctness with many scientists.

In the ensuing chapters, we shall show that these classical methods are in fact intellectually quite indefensible and do not deserve their social success. Indeed, we shall argue that the ideal of total objectivity is unattainable and that classical methods, which pose as guardians of that ideal, actually violate it at every turn; virtually none of those methods can be applied without a generous helping of personal judgment and arbitrary assumption.

[3] These characteristic italics were edited out of the posthumously published version of Lakatos's original mimeographed paper.

1.f The Plan of This Book

The thesis we propound in this book is that scientific reasoning is reasoning in accordance with the calculus of probabilities.

In Chapter 2 we introduce that calculus, along with the principal theorems that will later serve in an explanatory role with respect to scientific method; and we shall also in that chapter introduce the notion of a probability density, and describe some of the main probability distributions and associated theorems that will be needed later on.

We shall then, in Chapter 3, consider different interpretations of the probability calculus and introduce the notion of degrees of belief.

In Chapter 4 we examine how scientific theories may be confirmed or disconfirmed and look at characteristic patterns of inductive reasoning, particularly in relation to deterministic theories; we argue that these patterns are best understood as arguments in probability and that non-Bayesian approaches, by comparison, offer little or no insight into them.

Then, in Chapter 5, we turn to statistical hypotheses, where the philosophy of scientific inference has mostly been the preserve of statisticians. Far and away the most influential voice in statistics in recent times has been that of the classical statistician, and we shall therefore first give an account of the classical viewpoint and of its expression in the theories of significance tests and estimation.

In Chapter 6, we examine the practical areas of agricultural and clinical trials and argue that restrictions placed on such trials by classical principles of inference are largely unjustified and often harmful to their efficient and ethical conduct.

Another practical area where classical principles have been highly influential is in the study of the regression, or correlation of one physical parameter on others, and in Chapter 7 we argue that standard classical methods of regression analysis are also misconceived and harmful to scientific advance.

In Chapter 8, we outline the Bayesian approach to statistical inference and show its merits over its chief rivals. And finally, in the last chapter, we consider the commonest objections brought

against the Bayesian theory, principally the complaint that the subjectivism implicit in it is at variance with any proper understanding of the scientific process and of inductive inference. This, and the other objections we show are baseless.

CHAPTER 2
The Probability Calculus

2.a | The Axioms

The rules governing the assignment of probabilities, together with all the deductive consequences of those rules, are collectively called the *probability calculus*. Formally, the rules, or axioms, of the probability calculus assign non-negative real numbers (the probabilities), from among those between 0 and 1 inclusive, to a class of possible states of affairs, where these are represented under some appropriate manner of description. For the time being all that we shall assume about this class of representations, called the domain of discourse, or *domain* for short, is that it is closed under *conjoining* any two items with 'and', *disjoining* them with 'or', and *negating* any single item with 'not'. Thus if a and b represent possible states of affairs, so do respectively 'a and b', symbolised $a \,\&\, b$; 'a or b', symbolised $a \vee b$; and 'not-a', symbolised $\sim a$.

We shall allow for a certain amount of redundancy in the way the members of this possibility structure are characterised, just as we do in ordinary discourse. For example, '$\sim\sim a$' is just another, more complicated, way of saying a, and a and $\sim\sim a$ are logically equivalent. In general, if a and b are *logically equivalent* representations of any possible state we shall symbolise the fact by the notation $a \Leftrightarrow b$. It is useful (actually indispensable in the development of the formal theory) to consider as limiting cases those possible states of affairs which must necessarily occur, such as the state of its either raining or not raining, and those which necessarily cannot occur, such as its simultaneously raining and not raining (in a particular place). The symbolism $a \vee \sim a$ represents a necessary truth, and is itself is called a *logical truth*, while $a \,\&\, \sim a$ represents a necessary falsehood, and is called a *logical falsehood*, or *contradiction*. In what follows, t will be the generic

symbol of a logical truth and $\perp$ that of a contradiction. To any reader who has had exposure to an elementary logic course these concepts and the notation will be familiar as the formal basics of a propositional language, and for that reason we shall call these items, a, b, c, . . . and the compounds we can form from them, using the operations $\sim$, $\mathbf{v}$ and $\&$, *propositions*. The 'proposition' terminology is not ideal, but there is no better general-purpose term around to refer to classes of possible states of affairs, be they localised in spacetime or larger-scale types of possible world.

A word to the wise, that is, to those who have at some point consulted textbooks of probability, elementary or advanced. These texts frequently start off by defining a *probability-system*, which is a triple $(S, \Im, P)$, where P is a non-negative, real-valued function on $\Im$, which is called a *field* of subsets of S, where the latter is called variously the *class of elementary events*, *sample-space* or *possibility space*. That $\Im$ is a field of subsets of S means that it contains S itself, and is closed under the set-theoretic operations of *complementation with respect to S*, *union* and *intersection*. It follows that $\Im$ contains $\varnothing$, the empty set, since this is the complement of S with respect to itself. We can relate this to our own rather (in fact deliberately) informal treatment as follows. $\Im$ corresponds to our domain of propositions (referring to a class of possible states of affairs here represented by S), with negation represented by relative complement, conjunction by intersection, and disjunction by union. The only significant difference is that the set-theoretic formalism is purely *extensional*: there is no room for equivalent yet distinct descriptions of the same events in S. Thus, for example, S is the single extension of all the logically true propositions like $a \mathbf{v} \sim a$, $\sim(a \& \sim a)$, and so forth), and $\varnothing$ the single extension of all logical falsehoods. By writing t and $\perp$ as generic logical truths and falsehoods we are in effect performing notationally the same collapsing operation as is achieved by going set-theoretical.

A word to the very wise. Sometimes the probability function is said to be defined on a *Boolean algebra*, or algebra for short. A celebrated mathematical result lies behind this terminology, namely Stone's Theorem that every Boolean algebra is isomorphic to a field of sets. Thus we can talk of an algebra of sets, implicitly referring to the unique algebra isomorphic to the given

field. Also, the propositional operations of conjunction and disjunction are often symbolised using the Boolean-algebraic symbols for meet and join, $\wedge$ and $\vee$. The reason for this is that if we identify logically equivalent elements of a propositional language we also obtain a Boolean algebra, the so-called *Lindenbaum algebra* of the language. Sometimes, for this reason, people speak of an algebra of propositions. Strictly speaking, however, the elements of a propositional language are not isomorphic to a Boolean algebra, merely homomorphic, because the mapping is only many-one from the propositions to corresponding elements of the algebra (all logical truths map to the unique maximal element $\mathbf{1}$ of the algebra, and all logical falsehoods map to the unique least element $\mathbf{0}$, and in general all equivalent propositions map to the same member of the algebra; the reader might like to check that the algebra determined by one propositional variable has four members, that generated by two has sixteen, and that generated by n has 2 raised to the power 2^n members).

So much for peripheral technicalities. In what follows we shall regard probabilities as defined on domains of propositions closed under negation, conjunction, and disjunction, with the probability function on a particular domain denoted by P, and $P(a)$ read as 'the probability of a'. This brings us to the question of what $P(a)$ actually means. A remarkable fact about the probability calculus, discovered two hundred years ago, is that such statements can be endowed with two quite distinct types of meaning. One refers to the way the world is structured, and in particular the way it appears to endow certain types of *stochastic* (chance-like or random) experiment with a disposition to deliver outcomes in ways which betray marked large-scale regularities. Here the probabilities are objective, numerical measures of these regularities, evaluated empirically by the long-run relative frequencies of the corresponding outcomes. On the alternative interpretation the meaning of $P(a)$ is *epistemic* in character, and indicates something like the degree to which it is felt some assumed body of background knowledge renders the truth of a more or less likely, where a might be anything from a prediction about the next toss of a particular coin to a statement of the theory of General Relativity. These senses of $P(a)$ are not entirely unrelated. Knowing the objective probability of getting heads with a particular coin should, it seems

reasonable to believe, also tell you how likely it is that the next toss of the coin will yield a head.

We shall investigate these interpretative issues in more detail later. The task now is to get a feel for the formal principles of the probability calculus, and in particular see what the fundamental postulates are and discover some useful consequences of them. The fundamental postulates, known as the probability axioms, are just four in number:

(1) $P(a) \geq 0$ for all a in the domain of P.

(2) $P(t) = 1$.

(3) $P(a \vee b) = P(a) + P(b)$ if a and b are mutually inconsistent; that is, if $a \,\&\, b \Leftrightarrow \bot$.

(1)–**(3)** above suffice to generate that part of the probability calculus dealing with so-called *absolute* or *unconditional probabilities*. But a good deal of what follows will be concerned with probability functions of two variables, unlike P above which is a function of only one. These two-place probability functions are called *conditional probabilities*, and the conditional probability of a given b is written $P(a|b)$. There is a systematic connection between conditional and unconditional probabilities, however, and it is expressed in our fourth axiom:

(4) $P(a|b) = \dfrac{P(a \,\&\, b)}{P(b)}$ where $P(b) \neq 0$.

Many authors take $P(a|b)$ actually to be defined by **(4)**. We prefer to regard **(4)** as a postulate on a par with **(1)**–**(3)**. The reason for this is that in some interpretations of the calculus, independent meanings are given to conditional and unconditional probabilities, which means that in those **(4)** cannot be true simply by definition.

2.b Useful Theorems of the Calculus

The first result states the well-known fact that the probability of a proposition and that of its negation sum to 1:

(5) $P(\sim a) = 1 - P(a)$

Proof:

a entails $\sim\sim a$. Hence by **(3)** $P(a \vee a) = P(a) + P(\sim a)$. But by **(2)** $P(a \vee \sim a) = 1$, whence **(5)**.

Next, it is simple to show that contradictions have zero probability:

(6) $P(\bot) = 0$.

Proof:

$\sim\bot$ is a logical truth. Hence $P(\sim\bot) = 1$ and by (5) $P(\bot) = 0$.

Our next result states that equivalent sentences have the same probability:

(7) If $a \Leftrightarrow b$ then $P(a) = P(b)$.

Proof:

First, note that $a \vee \sim b$ is a logical truth if $a \Leftrightarrow b$.
Assume that $a \Leftrightarrow b$. Then $P(a \vee \sim b) = 1$. Also if $a \Leftrightarrow b$ then a entails $\sim\sim b$ so $P(a \vee \sim b) = P(a) + P(\sim b)$.
But by **(5)** $P(\sim b) = 1 - P(b)$, whence $P(a) = P(b)$.

We can now prove the important property of probability functions that they respect the entailment relation; to be precise, the probability of any consequence of a is at least as great as that of a itself:

(8) If a entails b then $P(a) \leq P(b)$.

Proof:

If a entails b then $[a \vee (b \,\&\, \sim a)] \Leftrightarrow b$. Hence by **(7)** $P(b) = P[a \vee (b \,\&\, \sim a)]$. But a entails $\sim(b \,\&\, \sim a)$ and so $P[a \vee (b \,\&\, \sim a)] = P(a) + P(b \,\&\, \sim a)$. Hence $P(b) = P(a) + P(b \,\&\, \sim a)$. But by **(1)** $P(b \,\&\, \sim a) \geq 0$, and so $P(a) \leq P(b)$.

From **(8)** it follows that probabilities are numbers between 0 and 1 inclusive:

(9) $0 \leq P(a) \leq 1$, for all a in the domain of P.

Proof:

By axiom 1, $P(a) \geq 0$, and since a entails t, where t is a logical truth, we have by **(8)** that $P(a) \geq P(t) = 1$.

We shall now demonstrate the general (finite) additivity condition:

(10) Suppose a_i entails $\sim a_j$, where $1 \leq i < j \leq n$. Then $P(a_1 \vee \ldots \vee a_n) = P(a_1) + \ldots + P(a_n)$.

Proof:

$P(a_1 \vee \ldots \vee a_n) = P[(a_1 \vee \ldots \vee a_{n-1}) \vee a_n]$, assuming that $n > 1$; if not the result is obviously trivial. But since a_i entails $\sim a_j$, for all $i \neq j$, it follows that $(a_1 \vee \ldots \vee a_{n-1})$ entails $\sim a_n$, and hence $P(a_1 \vee \ldots \vee a_n) = P(a_1 \vee \ldots \vee a_{n-1}) + P(a_n)$. Now simply repeat this for the remaining $a_1, \ldots, a_{n-1}$ and we have **(10)**. (This is essentially a proof by mathematical induction.)

Corollary. If $a_1 \vee \ldots \vee a_n$ is a logical truth, and a_i entails $\sim a_j$ for $i \neq j$, then $1 = P(a_1) + \ldots + P(a_n)$.

Our next result is often called the 'theorem of total probability'.

(11) If $P(a_1 \vee \ldots \vee a_n) = 1$, and a_i entails $\sim a_j$ for $i \neq j$, then $P(b) = P(b \,\&\, a_1) + \ldots + P(b \,\&\, a_n)$, for any proposition b.

Proof:

b entails $(b \,\&\, a_1) \vee \ldots \vee (b \,\&\, a_n) \vee [b \,\&\, \sim(a_1 \vee \ldots \vee a_n)]$. Furthermore, all the disjuncts on the right-hand side are mutually exclusive. Let $a = a_1 \vee \ldots \vee a_n$. Hence by **(10)** we have that $P(b) = P(b \,\&\, a_1) + \ldots + P(b \,\&\, a_n) + P(b \,\&\, \sim a)$. But $P(b \,\&\, \sim a) \leq P(\sim a)$, by **(8)**, and $P(\sim a) = 1 - P(a) = 1 - 1 = 0$. Hence $P(b \,\&\, \sim a) = 0$ and **(11)** follows.

Corollary 1. If $a_1 \vee \ldots \vee a_n$ is a logical truth, and a_i entails $\sim a_j$ for $i \neq j$, then $P(b) = \Sigma P(b \,\&\, a_i)$.

Corollary 2. $P(b) = P(b \mid c)\, P(c) + P(b \mid \sim c)\, P(\sim c)$, for any c such that $P(c) > 0$.

Another useful consequence of **(11)** is the following:

(12) If $P(a_1 \vee \ldots \vee a_n) = 1$ and a_i entails $\sim a_j$ for $i \neq j$, and $P(a_i) > 0$, then for any b, $P(b) = P(b \mid a_1)P(a_1) + \ldots + P(b \mid a_n)\, P(a_n)$.

Proof:

A direct application of **(4)** to **(11)**.

(12) itself can be generalized to:

If $P(a_1 \vee \ldots \vee a_n) = 1$ and $P(a_i \,\&\, a_j) = 0$ for all $i = j$, and $P(a_i) > 0$, then for any b, $P(b) = P(b \mid a_1)\, P(a_1) + \ldots + P(b \mid a_n)P(a_n)$.

We shall now develop some of the important properties of the function $P(a \mid b)$. We start by letting b be some fixed proposition such that $P(b) > 0$ and defining the function $Q(a)$ of one variable to be equal to $P(a \mid b)$, for all a.

Now define 'a is a logical truth modulo b' simply to mean 'b entails a' (for then a and t are equivalent given b), and 'a and c are exclusive modulo b' to mean '$b \,\&\, a$ entails $\sim c$'; then

(13) $Q(a) = 1$ if a is a logical truth modulo b; and the corollary

(14) $Q(b) = 1$;

(15) $Q(a \vee c) = Q(a) + Q(c)$, if a and c are exclusive modulo b.

Now let $Q'(a) = P(a \mid c)$, where $P(c) > 0$; in other words, Q' is obtained from P by fixing c as the conditioning statement, just as Q was obtained by fixing b. Since Q and Q' are probability functions on the same domain, we shall assume that axiom 4 also holds for them: that is, $Q(a \mid d) = \dfrac{Q(a \,\&\, d)}{Q(d)}$ where $Q(d) > 0$, and similarly for Q'. We can now state an interesting and important invariance result:

(16) $Q(a \mid c) = Q'(a \mid b)$.

Proof:

$$Q(a \,\&\, c) = \frac{Q(a \,\&\, c)}{Q(c)} \quad \frac{P(a \,\&\, b \mid c)}{P(c \mid b)} \quad \frac{P(a \,\&\, b \,\&\, c)}{P(b \,\&\, c)} =$$

$$\frac{P(a \,\&\, b \mid c)}{P(b \mid c)} \quad \frac{Q'(a \,\&\, b)}{Q'(b)} = Q(a \mid b).$$

Corollary. $Q(a \mid c) = P(a \mid b \,\&\, c) = Q'(a \mid b)$.

(16) and its corollary say that successively conditioning P on b and then on c gives the same result as if P were conditioned first on c and then on b, and the same result as if P were simultaneously conditioned on $b \,\&\, c$.

(17) If h entails e and $P(h) > 0$ and $P(e) < 1$, then $P(h \mid e) > P(h)$.

This is a very easy result to prove (we leave it as an exercise), but it is of fundamental importance to the interpretation of the probability calculus as a logic of inductive inference. It is for this reason that we employ the letters h and e; in the inductive interpretation of probability h will be some hypothesis and e some evidence. ***(17)*** *then states that if* h *predicts* e *then the occurrence of* e *will, if the conditions of* ***(17)*** *are satisfied, raise the probability of* h.

(17) is just one of the results that exhibit the truly inductive nature of probabilistic reasoning. It is not the only one, and more celebrated are those that go under the name of *Bayes's Theorems.* These theorems are named after the eighteenth-century English clergyman Thomas Bayes. Although Bayes, in a posthumously published and justly celebrated *Memoir* to the Royal Society of London (1763), derived the first form of the theorem named after him, the second is due to the great French mathematician Laplace.

Bayes's Theorem (First Form)

(18) $P(h \mid e) = \dfrac{P(e \mid h)\, P(h)}{P(e)}$, where $P(h), P(e) > 0$.

Proof:

$$P(h \mid e) = \frac{P(h \,\&\, e)}{P(e)} \quad \frac{P(e \mid h)\, P(h)}{P(e)}$$

Again we use the letters h and e, standing for hypothesis and evidence. This form of Bayes's Theorem states that the probability of the hypothesis conditional on the evidence (or the *posterior probability* of the hypothesis) is equal to the probability of the data conditional on the hypothesis (or the *likelihood* of the hypothesis) times the probability (the so-called *prior probability)* of the hypothesis, all divided by the probability of the data.

Bayes's Theorem (Second Form)

(19) *If* $P(h_1 \vee \ldots \vee h_n) = 1$ and h_i entails $\sim h_j$ *for* $i \neq j$ and $P(h_i), P(e) > 0$ then

$$P(h_k \mid e) = \frac{P(e \mid h_k)\, P(h_k)}{\Sigma P(e \mid h_i)\, P(h_i)}$$

Corollary. If $h_1 \vee \ldots \vee h_n$ is a logical truth, then if $P(e)$, $P(h_i) > 0$ and h entails $\sim h_j$ for $i \neq j$, then

$$P(h_k \mid e) = \frac{P(e \mid h_k)\, P(h_k)}{\Sigma P(e \mid h_i)\, P(h_i)}$$

Bayes's Theorem (Third Form)

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

From the point of view of inductive inference, this is one of the most important forms of Bayes's Theorem. For, since $P(\sim h) = 1 - P(h)$, it says that $P(h \mid e) = f\left(P(h), \dfrac{P(e \mid \sim h)}{P(e \mid h)}\right)$ where f is an increasing function of the prior probability $P(h)$ of h and a decreasing function of the *likelihood ratio* $\dfrac{P(e \mid \sim h)}{P(e \mid h)}$. In other words, for

a given value of the likelihood ratio, the posterior probability of h increases with its prior, while for a given value of the prior, the posterior probability of h is the greater, the less probable e is relative to $\sim h$ than to h.

2.c | Discussion

Despite their seemingly abstract appearance, implicit in axioms **(1)**–**(4)** is some very interesting, significant and sometimes surprising information, and a good deal of this book will be taken up with making it explicit and explaining why it is significant.

To whet the appetite, consider the following apparently simple problem, known as the Harvard Medical School Test (Casscells, Schoenberger, and Grayboys 1978), so called because it was given as a problem to students and staff at Harvard Medical School, whose responses we shall come to shortly.[1] A diagnostic test for a disease, D, has two outcomes 'positive' and 'negative' (supposedly indicating the presence and absence of D respectively). The test is a fairly sensitive one: its chance of giving a false negative outcome (showing 'negative' when the subject has D) is equal to 0, and its chance of giving a false positive outcome (showing 'positive' when the subject does not have D) is small: let us suppose it is equal to 5%. Suppose the incidence of the disease is very low, say one in one thousand in the population. A randomly selected person is given the test and shows a positive outcome. What is the chance they have D?

One might reason intuitively as follows. They have tested positive. The chance of testing positive and not having D would be only one in twenty. So the chance of having D given a positive result should be around nineteen twentieths, that is, 95%. This is the answer given by the majority of the respondents too. It is wrong; very wrong in fact: the correct answer is less than two in one hundred! Let us see why.

Firstly, anyone who answered 95% should have been suspicious that a piece of information given in the problem was not used, namely the incidence of D in the population. In fact, that information is highly relevant, because the correct calculation

[1] The discussion here follows Howson 2000, Chapter 3.

cannot be performed without it, as we now show. We can represent the false negative and false positive chances formally as conditional probabilities $P(\sim e \mid h) = 0$ and $P(e \mid \sim h) = 0.05$ respectively, where h is 'the subject has D' and e is 'the outcome is positive'. This means that our target probability, the chance that the subject has D given that they tested positive, is $P(h \mid e)$, which we have to evaluate. Since the subject was chosen randomly it seems reasonable to equate $P(h)$, the absolute probability of them having D, to 0.001, the incidence of D in the population. By **(5)** in section **b** we infer that $P(e \mid h) = 1$, and that $P(\sim h) = 0.999$. We can now plug these numbers into Bayes's Theorem in the form **(20)** in **b**, and with a little arithmetic we deduce that $P(h \mid e) = 0.0196$, that is, slightly less than 2%.

Gigerenzer (1991) has argued that the correct answer is more naturally and easily found from the data of the problem by translating the fractional chances into whole-number frequencies within some actual population of 1,000 people in which one individual has D, and that the diagnosis of why most people initially get the wrong answer, like the Harvard respondents, is due to the fact that the data would originally have been obtained in the form of such frequencies, and then been processed into chance or probability language which the human mind finds unfamiliar and unintuitive. Thus, in the Gigerenzer-prescribed format, we are looking to find the frequency of D-sufferers in the subpopulation of those who test positive. Well, since the false negative rate is zero, the one person having D should test positive, while the false negative rate implies that, to the nearest whole number, 50 of the 999 who don't have D will also test positive. Hence 51 test positive in total, of whom 1 by assumption has D. Hence the correct answer is now easily seen to be approximately 1 in 51, without the dubious aid of recondite and unintelligible formulas.

Caveat emptor![2] When something is more difficult than it apparently needs to be, there is usually some good reason, and there is a compelling reason why the Gigerenzer mode of reasoning is not to be recommended: it is invalid! As we shall see later, there is no direct connection between frequencies in finite samples and probabilities. One cannot infer directly anything about

[2] Buyer beware!

frequencies in finite samples from statements about a probability distribution, nor, conversely, can one infer anything directly about the latter from frequencies in finite samples. In particular, one is certainly not justified in translating a 5% chance of e conditional on $\sim h$ into the statement that in a sample of 999, 50 will test positive, and even less can one, say, translate a zero chance of e conditional on h into the statement that a single individual with D will test positive. As we shall also see later, the most that can be asserted is that *with a high probability* in a *big enough* sample the observed frequency will lie *within a given neighbourhood* of the chance. How we compute those neighbourhoods is the task of statistics, and we shall discuss it again in Chapters 5 and 8.

It is instructive to reflect a little on the significance of the probability-calculus computation we have just performed. It shows that the criteria of low false-positive and false-negative rates *by themselves* tell you nothing about how reliable a positive outcome is in any given case: an additional piece of information is required, namely the incidence of the disease in the population. The background incidence also goes by the name of 'the base rate', and thinking that valid inferences can be drawn just from the knowledge of false positive and negative rates has come to be called the 'base-rate fallacy'. As we see, if the base-rate is sufficiently low, a positive outcome in the Harvard Test is consistent with a very small chance of the subject having the disease, a fact which has profound practical implications: think of costly and possibly unpleasant follow-up investigations being recommended after a positive result for some very rare disease. The Harvard Test is nevertheless a challenge to the average person's intuition, which is actually rather poor when it come to even quite elementary statistical thinking. Translating into frequency-language, we see that even if it can be guaranteed that the null hypothesis (that the subject does not have the disease) will be rejected only very infrequently on the basis of an incorrect (positive) result, this is nevertheless consistent with almost all those rejections being incorrect, a fact that is intuitively rather surprising—which is of course why the base-rate fallacy is so entrenched.

But there is another, more profound, lesson to be drawn. We said that there are two quite distinct types of probability, both obeying the same formal laws **(1)**–**(4)** above, one having to do

with the tendency, or *objective probability*, of some procedure to produce any given outcome at any given trial, and the other with our uncertainty about unknown truth-values, and which we called *epistemic* probability, since it is to do with our knowledge, or lack of it. Since both these interpretations obey the same formal laws (we shall prove this later), it follows that *every formally valid argument involving one translates into a formally valid argument involving the other.*

This fact is of profound significance. Suppose h and e in the Harvard Test calculation had denoted some scientific theory under scrutiny and a piece of experimental evidence respectively, and that the probability function P is of the epistemic variety denoting something we can call 'degree of certainty'. We can infer that even if e had been generated by an experiment in which e is predicted by h but every unlikely were h to be false, that would still *by itself* give us no warrant to conclude anything about the degree of certainty we are entitled to repose in h[3]. To do that we need to plug in a value for $P(h)$, the prior probability of h. That does not means that you have to be able to compute $P(h)$ according to some uniform recipe; it merely means that in general you cannot make an inference ending with a value for $P(h \mid e)$ without putting some value on $P(h)$, or at any rate restricting it within certain bounds (though this is not always true, especially where there is a lot of experimental data where, as we shall see, the posterior probability can become almost independent of the prior).

The lessons of the Harvard Medical School Test now have a much more general methodological applicability. The results can be important and striking. Here are two examples. The first concerns what has been a major tool of statistical inference, *significance*

[3] That it does is implicit in the so-called Neyman-Pearson theory of statistical testing which we shall discuss later in some detail. And compare Mayo: if e 'fits' h [is to be expected on the basis of h] and there is a very small chance that the test procedure 'would yield so good a fit if h is false', then 'e should be taken as good grounds for h to the extent that h has passed a severe test with e' (1996, p.177; we have changed her upper case e and h to lower case). Mayo responds to the Harvard Medical School Test example in Mayo 1977, but at no point does she explain satisfactorily how obtaining an outcome which gives one less than a 2% chance of having the disease can possibly constitute 'good grounds' for the hypothesis that one has it.

testing, a topic we shall discuss in detail in Chapter 5. A Neyman-Pearson significance test is a type of so-called likelihood ratio test, where a region in the range of a test variable is deemed a rejection region depending on the value of a likelihood ratio on the boundary. This is determined in such a way that the probabilities of (a) the hypothesis being rejected if it is true, and (b) its being accepted if it is false, are kept to a minimum (the extent to which this is achievable will be discussed in Chapter 5). But these probabilities (strictly, probability-densities, but that does not affect the point) are, in effect, just the chances of a false negative and a false positive, and as we saw so graphically in the Harvard Medical School Test, finding an outcome in such a region *conveys no information whatever by itself* about the chance of the hypothesis under test being true.

The second example concerns the grand-sounding topic of *scientific realism*, the doctrine that we are justified in inferring to at least the approximate truth of a scientific theory T if certain conditions are met. These conditions are that the experimental data are exceptionally unlikely to have been observed if T is false, but quite likely if it is true. The argument, the so-called *No Miracles argument*, for the inference to the approximate truth of T is that if T is not approximately true then the agreement between T and the data are too miraculous to be due to chance (the use of the word 'miraculous', whence the name of the argument, was due to Putnam 1975). Again, we see essentially the same fallacious inference based on a small false positive rate and a small false negative rate as was committed by the respondents to the Harvard Test. However much we want to believe in the approximate truth of theories like quantum electrodynamics or General Relativity, both of which produce to order predictions correct to better than one part in a billion, the No Miracles argument is not the argument to justify such belief (a more extended discussion is in Howson 2000, Chapter 3).

2.d Countable Additivity

Before we leave this general discussion we should say something about a further axiom that is widely adopted in textbooks of mathematical probability: the *axiom of countable additivity*. This says that if $a_1, a_2, a_3, \ldots$ are a countably infinite family (this just means that they can be enumerated by the integers 1, 2, 3, . . .) of mutually inconsistent propositions in the domain of P and the statement 'One of the a_i is true' is also included in the domain of P then the probability of the latter is equal to the sum of the $P(a_i)$. Kolmogorov included a statement equivalent to it, his 'axiom of continuity', together with axioms **(1)**–**(4)** in his celebrated monograph (1950) as the foundational axioms of probability (except that he called **(4)** the 'definition' of conditional probability), and also required the domain of P to be closed not only under finite disjunctions (now *unions*, since the elements of the domain are now sets) but also countable ones, thus making it what is called a σ-field, or σ-algebra. These stipulations made probability a branch of the very powerful mathematical theory of *measure*, and the measure-theoretic framework has since become the paradigm for mathematical probability.

Mathematical considerations have undoubtedly been uppermost in this decision: the axiom of countable additivity is required for the strongest versions of the limit theorems of probability (characteristically prefaced by 'almost certainly', or 'with probability one', these locutions being taken to be synonymous); also the theory of random variables and distributions, particularly conditional distributions, receives a very smooth development if it is included. But we believe that the axioms we adopt should be driven by what logicians call 'soundness' considerations: their consequences should be *true* of whatever interpretation we wish to give them. And the brute fact is that for each of the principal interpretations of the probability calculus, the chance and the epistemic interpretation, not only are there no compelling grounds for thinking the countable additivity axiom always true but on the contrary there are good reasons to think it sometimes *false*.

The fact is that if we measure chances, or tendencies, by limiting relative frequencies (see Chapter 3) then we certainly have no reason to assume the axiom, since limiting relative frequencies, unlike finite frequencies in fixed-length samples, do not always obey it: in particular, if each of a countable infinity of exclusive and exhaustive possible outcomes tends to occur only finitely many times then its limiting relative frequency is zero,

while that of the disjunction is 1. As for the epistemic interpretation, as de Finetti pointed out (1972, p. 86), it may be perfectly reasonable (given suitable background information) to put a zero probability on each member of an exhaustive countably infinite partition of the total range of possibilities, but to do so contradicts the axiom since the probability of the total range is always 1. To satisfy the axiom of countable addivity the only permissible distribution of probabilities over a countable partition is one whose values form a sufficiently quickly converging sequence: for example, 1/2, 1/4, 1/8, . . . , and so forth. In other words, only very strongly *skewed* distributions are ever permitted over countably infinite partitions!

In both case, for chances and epistemic probabilities, therefore, there are cases where we might well want to assign equal probabilities to each of a countable infinity of exclusive and exhaustive outcomes, which we can do consistently if countable additivity is not required (but they must receive the uniform value 0), but would be prevented from doing so by the principle of countable additivity. It seems wrong in principle that an apparently gratuitous mathematical rule should force one to adopt instead a highly biased distribution. Not only that: a range of apparently very impressive convergence results, known in the literature as Bayesian convergence-of-opinion theorems, appear to show that under very general conditions indeed one's posterior probabilities will converge on the truth with probability one, where the truth in question is that of a hypothesis definable in a σ-field of subsets of an infinite product space (see, for example, Halmos 1950, p. 213, Theorem B). In other words, merely to be a consistent probabilistic reasoner appears to commit one to the belief that one's posterior probability of a hypothesis about an infinite sequence of possible data values will converge on certainty with increasing evidence. Pure probability theory, which we shall be claiming is no more than a type of logic, *as empty of specific content as deductive logic*, appears to be all that is needed to solve the notorious problem of induction!

If this sounds a bit too good to be true, it is: these results all turn out to require the principle of countable additivity for their proof, and exploit in some way or other the concentration of probability over a sufficiently large initial segment of a countably infinite partition demanded by the principle. To take a simple example from Kelly 1996, p. 323: suppose h says that a data source which can emit 0 or 1 emits only 1s on repeated trials, and that $P(h) > 0$. So h is false if and only if a 0 occurs at some point in an indefinitely extended sample. The propositions a_n saying that a 0 occurs first at the nth repetition are a countably infinite disjoint family, and the probability of the statement that at least one of the a_i is true, given the falsity of h, must be 1. So given the front-end skewedness prescribed by the axiom of countable additivity, the probability that h is false will be mostly concentrated on some finite disjunction $a_1 \mathbf{v} \ldots \mathbf{v} a_n$. It is left to the reader to show, as an easy exercise in Bayes's Theorem in the form **(20)**, section **b** above, that the probability that h is true, given a sufficiently long unbroken run of 1s, is very close to 1.

There is (much) more to be said on this subject, but for further discussion the reader is encouraged to consult de Finetti 1872, Kelly 1996, pp. 321–330, and Bartha 2004. Kelly's excellent book is particularly recommended for its illuminating discussion of the roles played not only by countable additivity but also (and non-negligibly) by the topological complexity of the hypotheses in probabilistic convergence-to-the-truth results.

2.e | Random Variables

In many applications the statements in the domain of P are those ascribing values, or intervals of values, to random variables. Such statements are the typical mode of description in statistics. For example, suppose we are conducting simultaneous measurements of individuals' heights and weights in pounds and metres. Formally, the set S of relevant possible outcomes will consist of all pairs $s = (x, y)$ of non-negative real numbers up to some big enough number for each of x and y, height and weight respectively (measuring down to a real number is of course practically impossible, but that is why this is an idealisation).

We can define two functions X and Y on S such that $X(x, y) = x$ and $Y(x, y) = y$. X and Y are examples of *random variables*: X picks out the height dimension, and Y the weight dimension of the various joint possibilities. In textbooks of mathematical probabil-

ity or statistics, a typical formula might be $P(X > x)$. What does this mean? The answer, perhaps not surprisingly, will depend on which of the two interpretations of P mentioned earlier is in play. On the chance interpretation, $P(X > x)$ will signify *the tendency of the randomising procedure to generate a pair of observations* (x', y') *satisfying the condition that* $x' > x$, and this tendency, as we observed, will be evaluated by inspecting the frequency with which it does generate such pairs.

On the other, epistemic, interpretation, $P(X > x)$ will signify a degree of uncertainty about some specific event signified by the same inequality formula $X > x$. For example, suppose that we are told that someone has been selected, possibly *but not necessarily* by a randomising procedure, but we know nothing about their identity. We are for whatever reason interested in the magnitude of their height, and entertain a range of conjectures about it, assigning uncertainty-probabilities to them. One such conjecture might be 'The height of the person selected exceeds x metres', and $P(X > x)$ now symbolises the degree of certainty attached to it.

This second reading shows that 'random variable' does not have to refer to a random procedure: there, it was just a way of describing the various possibilities determined by the parameters of some application. Indeed, not only do random variables have nothing necessarily to do with randomness, but *they are not variables either*: as we saw above, X, Y, etc. are not variables at all but, since they take different values depending on which particular possibilities are instantiated, *functions* on an appropriate possibility-space (in the full measure-theoretic treatment, their technical name is *measurable functions*).

2.f Distributions

Statements of the form '$X < x$', '$X \leq x$', play a fundamental role in mathematical statistics. Clearly, the probability of any such statement (assuming that they are all in the domain of the probability function) will vary with the choice of the real number x; it follows that this probability is a function $F(x)$, the so-called *distribution function*, of the random variable X. Thus, where P is the

probability measure concerned, the value of $F(x)$ is defined to be equal, for all x to $P(X \leq x)$ (although F depends therefore also on X and P, these are normally apparent from the context and F is usually written as a function of x only). Some immediate consequences of the definition of $F(x)$ are that

(i) if $x_1 < x_2$ then $F(x_1) \leq F(x_2)$, and

(ii) $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$.

Distribution functions are not necessarily functions of one variable only. For example, we might wish to describe a possible eventuality in terms of the values taken by a number of random variables. Consider the 'experiment' which consists in noting the heights (X, say) and weights (Y) jointly of members of some human population. It is usually accepted as a fact that there is a joint (objective) probability distribution for the vector variable (X, Y), meaning that there is a probability distribution function $F(x, y) = P(X \leq x \,\&\, Y \leq y)$. Mathematically this situation is straightforwardly generalised to distribution functions of n variables.

2.g Probability Densities

It follows from (ii) that if $F(x)$ is differentiable at the point x, then the *probability density* at the point x is defined and is equal to $f(x) = \dfrac{dF(x)}{dx}$ in other words, if you divide the probability that X is in a given interval $(x, x + h)$ by the length h of that interval and let h tend to 0, then if F is differentiable, there is a probability density at the point x, which is equal to $f(x)$. If the density exists at every point in an interval, then the associated probability distribution of the random variable is said to be continuous in that interval. The simplest continuous distribution, and one which we shall refer to many times in the following pages, is the so-called *uniform distribution*. A random variable X is uniformly distributed in a closed interval I if it has a constant positive probability density at every point in I and zero density outside that interval.

Probability densities are of great importance in mathematical statistics—indeed, for many years the principal subject of research in that field was finding the forms of density functions of random variables obtained by transformations of other random variables. They are so important because many of the probability distributions in physics, demography, biology, and similar fields are continuous, or at any rate approximate continuous distributions. Few people believe, however, in the real—as opposed to the mathematical—existence of continuous distributions, regarding them as only idealisations of what in fact are discrete distributions.

Many of the famous distribution functions in statistics are identifiable only by means of their associated density functions; more precisely, those cumulative distribution functions have no representation other than as integrals of their associated density functions. Thus the famous *normal distributions* (these distributions, of fundamental importance in statistics, are uniquely determined by the values of two parameters, their mean and standard deviation, which we shall discuss shortly) have distribution functions characterised as the integrals of density functions.

Some terminology. Suppose X and Y are jointly distributed random variables with a continuous distribution function $F(X, Y)$ and density function $f(x, y)$. Then $F(X) = \int_{-\infty}^{\infty} f(x, y)dy$ is called the *marginal distribution* of X. The operation of obtaining marginal distributions by integration in this way is the continuous analogue of using the theorem of total probability to obtain the probability $P(a)$ of a by taking the sum $\Sigma P(a \,\&\, b_i)$. Indeed, if X and Y are discrete, then the marginal distribution for X is just the sum $P(X = x_i) = \Sigma_j P(X = x_i \,\&\, Y = y_j)$. The definitions are straightforwardly generalised to joint distributions of n variables.

2.h | Expected Values

The *expected value* of a function $g(X)$ of X is defined to be (where it exists) the probability-weighted average of the values of g. To take a simple example, suppose that g takes only finitely many values $g_1, \ldots, g_n$ with probabilities $a_1, \ldots, a_n$. Then the expected value $E(g)$ of g always exists and is equal to $\Sigma g_i a_i$. If X has a probability density function $f(x)$ and g is integrable, then $E(g) = \int_{-\infty}^{\infty} g(x)f(x)dx$ where the integral exists.

In most cases, functions of random variables are themselves random variables. For example, the sum of any n random variables is a random variable. This brings us to an important property of expectations: they are so-called *linear functionals*. In other words, if $X_1, \ldots, X_n$ are n random variables, then if the expectations exist for all the X_i, then, because expectations are either sums or limits of sums, so does the expected value of the sum $X = X_1 + \ldots + X_n$ and $E(X) = E(X_1) + \ldots + E(X_n)$.

2.i | The Mean and Standard Deviation

Two quantities which crop up all the time in statistics are the mean and standard deviation of a random variable X. The *mean value* of X is the expected value $E(X)$ of X itself, where that expectation exists; it follows that the mean of X is simply the probability-weighted average of the values of X. The *variance* of X is the expected value of the function $(X - m)^2$, where that expectation exists. The *standard deviation* of X is the square root of the variance. The square root is taken because the standard deviation is intended as a characteristic measure of the spread of X away from the mean and so should be expressed in units of X. Thus, if we write s.d.(X) for the standard deviation of X, s.d.$(X) = \sqrt{E[(X - m)^2]}$, where the expectation exists. The qualification 'where the expectation exists' is important, for these expected values do not always exist, even for some well-known distributions. For example, if X has the Cauchy density $\frac{a}{\pi(a^2 + x^2)}$ then it has neither mean nor variance.

We have already mentioned the family of *normal distributions* and its fundamental importance in statistics. This importance derives from the facts that many of the variables encountered in nature are normally distributed and also that the sampling distributions of a great number of statistics tend to the normal as the size of the sample tends to infinity (a statistic is a numerical function of the observations, and hence a random variable). For the moment we shall confine the discussion to normal distributions of

one variable. Each member of this family of distributions is completely determined by two parameters, its mean μ and standard deviation σ. The normal distribution function itself is given by the integral over the values of the real variable t from $-\infty$ to x of the density we mentioned above, that is, by

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}$$

It is easily verified from the analytic expression for $F(x)$ that the parameters μ and σ are indeed the mean and standard deviation of X. The curve of the normal density is the familiar bell-shaped curve symmetrical about $x = \mu$ with the points $x = \mu \pm \sigma$ corresponding to the points of maximum slope of the curve (Figure 2.1). For these distributions the mean coincides with the *median*, the value of x such that the probability of the set $\{X < x\}$ is one half (these two points do not coincide for all other types of distribution, however). A fact we shall draw on later is that the interval on the x-axis determined by the distance of 1.96 standard deviations centred on the mean supports 95% of the area under the curve, and hence receives 95% of the total probability.

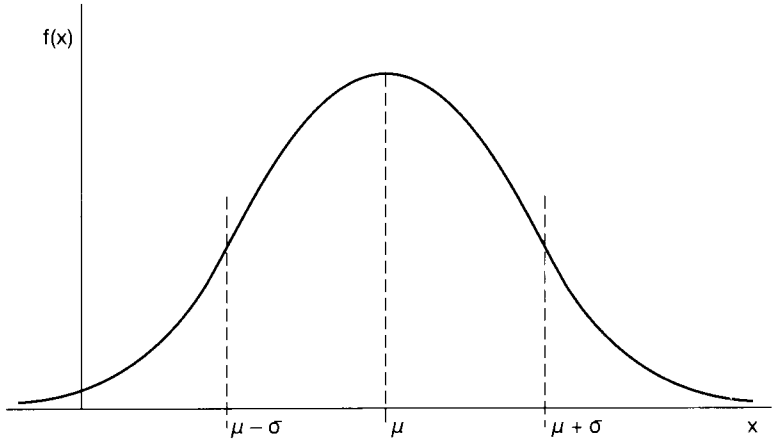


FIGURE 2.1

2.j | Probabilistic Independence

Two propositions h_1 and h_2 in the domain of P are said to be *probabilistically independent* (relative to some given probability measure P if and only if $P(h_1 \,\&\, h_2) = P(h_1)P(h_2)$. It follows immediately that, where $P(h_1)$ and $P(h_2)$ are both greater than zero, so that the conditional probabilities are defined, $P(h_1 \mid h_2) = P(h_1)$ and $P(h_2 \mid h_1) = P(h_2)$, just in case h_1 and h_2 are probabilistically independent.

Let us consider a simple example, which is also instructive in that it displays an interesting relationship between probabilistic independence and the so-called Classical Definition of probability. A repeatable experiment is determined by the conditions that a given coin is to be tossed twice and the resulting uppermost faces are to be noted in the sequence in which they occur. Suppose each of the four possible types of outcome—two heads, two tails, a head at the first throw and a tail at the second, a tail at the first throw and a head at the second—has the same probability, which of course must be one quarter. A convenient way of describing these outcomes is in terms of the values taken by two random variables X_1 and X_2, where X_1 is equal to 1 if the first toss yields a head and 0 if it is a tail, and X_2 is equal to 1 if the second toss yields a head and 0 if a tail.

According to the Classical Definition, or, as we shall call it, the Classical Theory of Probability, which we look at in the next chapter (and which should not be confused with the Classical Theory of Statistical Inference, which we shall also discuss), the probability of the sentence '$X_1 = 1$' is equal to the ratio of the number of those possible outcomes of the experiment which satisfy that sentence, divided by the total number, namely four, of possible outcomes. Thus, the probability of the sentence '$X_1 = 1$' is equal to 1/2, as is also, it is easy to check, the probability of each of the four sentences of the form '$X_i = X_i$', $i = 1$ or 2, $x_i = 0$ or 1. By the same Classical criterion, the probability of each of the four sentences '$X_1 = x_1$ & $X_2 = x_2$' is 1/4.

Hence

$$P(X_1 = x_1 \,\&\, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2)$$

and consequently the pairs of sentences '$X_1 = x_1$', '$X_2 = x_2$' are probabilistically independent.

The notion of probabilistic independence is generalised to n propositions as follows: $h_1, \ldots, h_n$ are said to be probabilistically independent (relative to the measure P) if and only if for every subset $h_{i1}, \ldots, h_{ik}$ of $h_1, \ldots, h_n$,

$$P(h_{i1} \,\&\, \ldots \,\&\, h_{ik}) = P(h_{i1}) \ldots P(h_{ik}).$$

It is easy to see, just as in the case of the pairs, that if any set of propositions is probabilistically independent, then the probability of any one of them being conditional on any of the others, where the conditional probabilities are defined, is the same as its unconditional probability. It is also not difficult to show (and it is, as we shall see shortly, important in the derivation of the binomial distribution) that if $h_1, \ldots, h_n$ are independent, then so are all the 2^n sets $\pm h_1, \ldots, \pm h_n$, where $+h$ is h and $-h$ is $\sim h$.

Any n random variables $X_1, \ldots, X_n$ are said to be independent if for all sets of intervals $I_1, \ldots, I_n$ of values of $X_1, \ldots, X_n$ respectively, the propositions $X_1 \in I_1, \ldots, X_n \in I_n$ are probabilistically independent. We have, in effect, already seen that the two random variables X_1 and X_2 in the example above are probabilistically independent. If we generalise that example to that of the coin's being tossed n times, and define the random variables $X_1, \ldots, X_n$ just as we defined X_1 and X_2, then again a consequence of applying the Classical 'definition' to this case is that $X_1, \ldots, X_n$ are probabilistically independent. It is also not difficult to show that a necessary and sufficient condition for any n random variables $X_1, \ldots, X_n$ to be independent is that

$$F(x_1, \ldots, x_n) = F(x_1) \ldots F(x_n)$$

where $F(x_1, \ldots, x_n)$ is the joint distribution function of the variables $X_1, \ldots, X_n$ and $F(x_i)$ is the marginal distribution of X_i. Similarly, if it exists, the joint density $f(x_1, \ldots, x_n)$ factors into the product of marginal densities $f(x_1) \ldots f(x_n)$ if the X_i are independent.

2.k | Conditional Distributions

According to the conditional probability axiom, axiom 4,

(1) $$P(X < x \mid y < Y < y + \delta y) = \frac{P(X < x \,\&\, y < Y \leq y + \delta y)}{P(y < Y \leq y + \delta y)}.$$

The left-hand side is an ordinary conditional probability. Note that if $F(x)$ has a density $f(x)$ at the point x, then $P(X = x) = 0$ at that point. We noted in the discussion of (4) that $P(a \mid b)$ is in general only defined if $P(b) > 0$. However, it is in certain cases possible for b to be such that $P(b) = 0$ and for $P(a \mid b)$ to take some definite value. Such cases are afforded where b is a sentence of the form $Y = y$ and there is a probability density $f(y)$ at that point. For then, if the joint density $f(x, y)$ also exists, then multiplying top and bottom in (1) by δy, we can see that as δy tends to 0, the right-hand side of that equation tends to the quantity

$$\int_{-\infty}^{x} \frac{f(u,y)du}{f(y)},$$

where $f(y)$ is the marginal density of y, which determines a distribution function for X, called the *conditional distribution function of* X with respect to the event $Y = y$. Thus in such cases there is a perfectly well-defined conditional probability

$$P(x_1 < X \leq x_2 \mid Y = y),$$

even though $P(Y = y) = 0$.

The quantity $\frac{f(x,y)}{f(y)}$ is the density function at the point $X = x$ of this conditional distribution (the point $Y = y$ being regarded now as a parameter), and is accordingly called the *conditional probability density of* X at x, relative to the event $Y = y$. It is of great importance in mathematical statistics and it is customarily denoted by the symbol $f(x \mid y)$. Analogues of **(18)** and **(19)**, the

two forms of Bayes's Theorem, are now easily obtained for densities: where the appropriate densities exist

$$f(x \mid y) = \frac{f(y \mid x)f(x)}{f(y)}$$

and

$$f(x \mid y) = \frac{f(y \mid x)f(x)}{\int_{-\infty}^{\infty} f(y \mid x)f(x)dx}.$$

2.l The Bivariate Normal

We can illustrate some of the abstract formal notions we have discussed above in the context of a very important multivariate distribution, the *bivariate normal distribution.* This distribution is, as its name implies, a distribution over two random variables, and it is determined by five parameters. The marginal distributions of the two variables X and Y are both themselves normal, with means μ_X, μ_Y and standard deviations σ_X, σ_Y. One more parameter, the correlation coefficient ρ, completely specifies the distribution. The bivariate density is given by

$$f(x,y) = \frac{e^{\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}.$$

This has the form of a more-or-less pointed, more-or-less elongated hump over the x, y plane, whose contours are ellipses with eccentricity (departure from circularity) determined by ρ. ρ lies between -1 and $+1$ inclusive. When $\rho = 0$, X and Y are *uncorrelated,* and the contour ellipses are circles. When ρ is either $+1$ or -1 the ellipses degenerate into straight lines. In this case all the probability is carried by a set of points of the form $y = ax + b$, for specified a and b, which will depend on the means and standard deviations of the marginal distributions. It follows that the conditional probability $P(X = x \mid Y = y)$ is 1 if $y = ax + b$, and 0 if not.

The conditional distributions obtained from bivariate (and more generally multivariate) normal distributions have great

importance in the area of statistics known as *regression analysis.* It is not difficult to show that the mean $\mu(X \mid y) = \int_{-\infty}^{\infty} xf(x \mid y)dx$ (or the sum where the conditional distribution is discrete) has the equation $\mu(X \mid y) = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y)$. In other words, the dependence of the mean on y is linear, with gradient , proportional to ρ, and this relationship defines what is called the regression of X on Y. The linear equation above implies the well-known phenomenon of *regression to the mean.* Suppose $\rho_X = \rho_Y$ and $\mu_X = \mu_Y = m$. Then $\mu(X \mid y) = m + \rho(y - m)$, which is the point located a proportion ρ of the distance between y and m. For example, suppose that people's heights are normally distributed and that Y is the average of the two parents' height and X is the offspring's height. Suppose also that the means and standard deviations of these two variables are the same and that $\rho = 1/2$. Then the mean value of the offspring's height is halfway between the common population mean and the two parents' average height. It is often said that results like this explain what we actually observe, but explaining exactly how parameters of probability distributions are linked to what we can observe turns out to be a hotly disputed subject, and it is one which will occupy a substantial part of the remainder of this book.

Let us leave that topic in abeyance, then, and end this brief outline of that part of the mathematical theory of probability which we shall have occasion to use, with the derivation and some discussion of the limiting properties of the first non-trivial random-variable distribution to be investigated thoroughly, and which has no less a fundamental place in statistics than the normal distribution, to which it is intimately related.

2.m The Binomial Distribution

This was the binomial distribution. It was through examining the properties of this distribution that the first great steps on the road to modern mathematical statistics were taken, by James Bernoulli, who proved (in *Ars Conjectandi,* published posthumously in 1713) the first of the limit theorems for sequences of independent random variables, the so-called *Weak Law of Large Numbers,* and

Abraham de Moivre, an eighteenth-century Huguenot mathematician settled in England, who proved that, in a sense we shall make clear shortly, the binomial distribution tends for large n to the normal. Although Bernoulli demonstrated his result algebraically, it follows, as we shall see, from de Moivre's limit theorem.

Suppose (i) X_i, $i = 1, \ldots, n$, are random variables which take two values only, which we shall label 0 and 1, and that the probability that each takes the value 1 is the same for all i, and equals p:

$$P(X_i = 1) = P(X_j = 1) = p.$$

Suppose also (ii) that the X_i are independent; that is,

$$P(X_1 = x_1 \,\&\, \ldots \,\&\, X_n = x_n) = P(X_1 = x_1) \times \ldots \times P(X_n = x_n),$$

where $x_i = 1$ or 0. In other words, the X_i are independent, identically distributed random variables. Let $Y_{(n)} = X_1 + \ldots + X_n$. Then for any r, $0 \leq r \leq n$,

$$\textbf{(2)} \quad P(Y_{(n)} = r) = {}^nC_r p^r (1-p)^{n-r}$$

since using the additivity property, the value of P is obtained by summing the probabilities of all conjunctions

$$X_1 = x_1 \,\&\, \ldots \,\&\, X_n = x_n,$$

where r of the x_i are ones and the remainder are zeros. There are ${}^nC_r$ of these, where ${}^nC_r$ is the number of ways of selecting r objects out of n, and is equal to $\dfrac{n!}{(n-r)!r!}$, where $n!$ is equal to $n(n-1)(n-2)\ldots 2.1$, and 0! is set equal to 1). By the independence and constant probability assumptions, the probability of each conjunct in the sum is $p^r(1-p)^{n-r}$, since $P(X_i = 0) = 1 - p$.

$Y_{(n)}$ is said to possess the *binomial distribution*. The mean of $Y_{(n)}$ is np, as can be easily seen from the facts that

$$E(X_1 + \ldots + X_n) = E(X_1) + \ldots + E(X_n)$$

and that

$$E(X_i) = p \cdot 1 + (1-p) \cdot 0 = p.$$

The squared standard deviation, or variance of $Y_{(n)}$, is

$$\begin{aligned} E(Y_{(n)} - np)^2 &= E(Y_{(n)}{}^2) + E(np)^2 - E(2Y_{(n)}np) \\ &= E(Y_{(n)}{}^2) + (np)^2 - 2npE(Y_{(n)}) \\ &= E(Y_{(n)}{}^2) - (np)^2. \end{aligned}$$

Now

$$\begin{aligned} E(Y_{(n)}{}^2) &= \Sigma(X^2{}_i) + \Sigma_{i \neq j} E(X_i X_j) \\ &= np + n(n-1)p^2. \end{aligned}$$

Hence

$$\text{s.d.}(Y_{(n)}) = \sqrt{np^2 + np} = \sqrt{np(1-p)}.$$

2.m | The Weak Law of Large Numbers

The significance of these expressions is apparent when n becomes very large. De Moivre showed that for large n, $Y_{(n)}$ is approximately normally distributed with mean np and standard deviation $\sqrt{np(1-p)}$ (the approximation is very close for quite moderate values of n). This implies that the so-called *standardised variable* $Z = \dfrac{(Y_{(n)} - np)}{\sqrt{np(1-p)}}$ is approximately normally distributed for large n, with mean 0 and standard deviation 1 (Z is called 'standardised' because it measures the distance of the relative frequency from its mean in units of the standard deviation). Hence

$$P(-k < Z < k) \approx \Phi(k) - \Phi(-k),$$

where Φ is the normal distribution function with zero mean and unit standard deviation. Hence

$$P(p - k\sqrt{\frac{pq}{n}} < \frac{Y}{n} < p + k\sqrt{\frac{pq}{n}}) \approx \Phi(k) - \Phi(-k),$$

where $q = 1 - p$. So, setting $\varepsilon = k\sqrt{\frac{pq}{n}}$,

$$P(p - \varepsilon < \frac{Y}{n} < p + \varepsilon) \approx \Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) - \Phi\left(-\varepsilon\sqrt{\frac{n}{pq}}\right)$$

Clearly, the right-hand side of this equation tends to 1, and we have obtained the *Weak Law of Large Numbers:*

$$P(|\frac{Y}{n} - p| < \varepsilon) \to 1, \text{ for all } \varepsilon > 0.$$

This is one of the most famous theorems in the history of mathematics. James Bernoulli proved it originally by purely combinatorial methods. It took him twenty years to prove, and he called it his "golden theorem". It is the first great result of the discipline now known as mathematical statistics and the forerunner of a host of other limit theorems of probability. Its significance outside mathematics lies in the fact that sequences of independent binomial random variables with constant probability, or Bernoulli sequences as they are called, are thought to model many types of sequence of repeated stochastic trials (the most familiar being tossing a coin n times and registering the sequence of heads and tails produced). What the theorem says is that for such sequences of trials the relative frequency of the particular character concerned, like heads in the example we have just mentioned, is with arbitrarily great probability going to be situated arbitrarily close to the parameter p.

The Weak Law, as stated above, is only one way of appreciating the significance of what happens as n increases. As we saw, it was obtained from the approximation

$$P(p - k\sqrt{\frac{pq}{n}} < \frac{Y}{n} < p + k\sqrt{\frac{pq}{n}}) \approx \Phi(k) - \Phi(-k),$$

where $q = 1 - p$, by replacing the variable bounds (depending on

$n) \pm k\sqrt{\frac{pq}{n}}$ by ε, and replacing k on the right-hand side by $\varepsilon\sqrt{\frac{n}{pq}}$.

The resulting equation is equivalent to the first. In other words, the Weak Law can be seen either as the statement that if we select some fixed interval of length 2ε centred on p, then in the limit as n increases, all the distribution will lie within that interval, or as the statement that if we first select any value between 0 and 1 and consider the interval centred on p which carries that value of the probability, then the endpoints of the interval move towards p as n increases, and in the limit coincide with p.

Another 'law of large numbers' seems even more emphatically to point to a connection between probabilities and frequencies in sequences of identically distributed, independent binomial random variables. This is the so-called Strong Law, which is usually stated as a result about actually infinite sequences of such variables: it asserts that with probability equal to 1, the limit of $Y_{(n)}/n$ exists (that is to say, the relative frequency of ones converges to some finite value) and is equal top.

So stated, the Strong Law requires for its proof the axiom of countable additivity, which we have cautioned against accepting as a general principle. Nevertheless, a 'strong enough' version of the Strong Law can be stated which does not assume countable additivity (the other 'strong' limit theorems of mathematical probability can usually be rephrased in a similar way): it says that for an infinite sequence $X_1, X_2, \ldots$ of $\{0,1\}$-valued random variables, if δ, ε are any positive numbers, however small, then there exists an n such that for all $m>n$ the probability that $Y_{(n+m)} - p$ is less than ε is greater than $1 - \delta$.

What this version of the Strong Law says is that the convergence of the $Y_{(n)}$ is *uniform* in the small probability. The Weak Law is weak in the sense that it merely says that the probability that the deviation of $Y_{(n)}$ from p is smaller than ε can be made arbitrarily close to 1 by taking n large enough; the Strong Law says that the probability that the deviation will become *and remain* smaller than ε can be made arbitrarily close to 1 by taking n large enough.

At any rate, throughout the eighteenth and nineteenth centuries people took these results to justify inferring, from the

observed relative frequency of some given character in long sequences of apparently *causally* independent trials, the approximate value of the postulated binomial probability. While such a practice may seem suggested by these theorems, it is not clear that it is in any way justified. While doubts were regularly voiced over the validity of this 'inversion', as it was called, of the theorem, the temptation to see in it a licence to infer to the value of p from 'large' samples persists, as we shall see in the next chapter, where we shall return to the discussion.

CHAPTER 3
The Laws of Probability

3.a Prologue: Frequency-Probability

We pointed out in 2.c that there are two main ways of interpreting the probability axioms. Throughout this book we shall be mainly concerned with one of them, an epistemic interpretation in which the probability function measures an agent's uncertainty. This interpretation is also called *Bayesian probability*. However, some discussion of the other interpretation of the axioms is unavoidable, because much of the application of Bayesian probability is to hypotheses about these other probabilities. There is a slight problem of terminology here, since there is no ready antonym to 'epistemic', but to emphasise the fact that these probabilities are supposed to characterise objective factual situations and set-ups we shall rest content with the not wholly satisfactory terminology of 'objective probabilities'.

What makes these important to science is that in a variety of contexts, from demography in the large to quantum mechanics in the small, they seem to be readily amenable to empirical measurement, at any rate in principle. For to say that an experimental arrangement has a stable objective probability of delivering any one of its possible outcomes is at least arguably to say that it has a characteristic *tendency* to do so, the magnitude of which is plausibly gauged in terms of the normalised, or relative, *frequency* with which the outcome in question is actually produced; or in other words, the number of times the outcome occurs in n trials, divided by n. Thus Pitowsky:

> The observational counterparts of the theoretical concept 'probability distribution' are the relative frequencies. In other words, as far as repeatable . . . events are concerned, probability is manifested in frequency. (1994, p. 98)

But not just frequency: fundamental to the identification of objective probability with relative frequency is *the long run*, for it is only in the long run, in general, that such frequencies behave as sufficiently stable features to be the counterparts of the postulated probability-distribution, and indeed to be objects of scientific enquiry at all. Moreover, there is a good deal of evidence that in suitable experimental contexts the relative frequency with which each of the various possible outcomes occurs settles down within a smaller and smaller characteristic interval as the number of observations increases. This is not very precise, however (how should the interval vary as a function of sample size?), and we shall follow von Mises (1939) in replacing the rather vague notion of a relative frequency 'settling down within an increasingly small interval' by the precise mathematical definition of a *limit*. Thus, where a describes a generic event A in the outcome-space of the experiment,[1] and $n(A)$ is the number of occurrences of A in n repetitions of the experiment, we define the measure of the probability (we shall often simply write 'probability' where the context shows that it is objective probability that is being referred to) of a to be the limit of $n(A)/n$ as n tends to infinity. This limit is to be regarded as a characteristic attribute of the experimental conditions themselves. Indeed, just as tendencies manifest themselves in frequencies of occurrence when the conditions are repeatedly instantiated, we can think of this postulated limiting relative frequency as an *exact* measure of the tendency of those conditions to deliver the outcome in question (this idea has of course been promoted under the name 'propensity theory of probability', by Popper (1959) and others; it was also the view of von Mises whose own frequency theory is usually, and incorrectly, regarded as not being of this type).

A pleasing feature of the limit definition is that limiting frequencies demonstrably satisfy the axioms of the probability calculus. This is very easy to show for unconditional probabilities, and we leave it as an exercise. For conditional probabilities the situation is only slightly more complicated, due to the fact that as yet conditional probabilities have not been defined in a frequency-context. But the definition is entirely natural and intuitive: where a and b are generic event-descriptors as above, we define $P(a|b)$ to be the long-run relative frequency of outcomes of type A among all those of type B. Hence $P(a|b) = \lim_{n\to\infty} n(A\&B)/n(B)$. It immediately follows that $P(a|b) = \lim_{\to\infty} [n(A\&B)/n \div n(B)/n]$ where $\lim n(B)$ exists. Hence $P(a|b) = P(a\&b)/P(b)$ where $P(b) > 0$.

[1] A generic event-description is one which states simply that the event occurs, without reference to any specific feature of the situation, its time, place, and so forth. So, if the experiment consists in tossing a given coin, A might be the event of landing heads, in which case a is the statement 'the coin lands heads'.

While the use of the limit definition allows us to regard objective probabilities as probabilities in the purely formal sense of the probability calculus, it has nevertheless elicited from positivistically-minded philosophers and scientists the objection that we can never in principle, not just in practice, observe the infinite n-limits. Indeed, we know that in fact (given certain plausible assumptions about the physical universe) *these limits do not exist*. For any physical apparatus would wear out or disappear long before n got to even moderately large values. So it would seem that no empirical sense can be given to the idea of a limit of relative frequencies. To this apparently rather strong objection another is frequently added, that defining chances in terms of limiting frequencies is anyway unnecessary: the mathematical theory of probability itself, in the form of the Strong Law of Large Numbers shows how the apparent convergence of the relative frequencies to a limit can be *explained* as a feature of long sequences of Bernoulli trials.[2]

Let us look into this claim. The 'strongest' form of the Strong Law refers to the set of possible outcomes w of an infinitely repeated experiment which generates at each repetition either a 0 or a 1. These possibilities are described by a sequence X_i of independent, identically distributed random variables such that $X_i(w)$ is equal to the ith coordinate of w (0 or 1) and $P(X_i = 1) = p$ for some fixed p and all i. This Strong Law says that with probability 1 the limit, as

[2] A variant of the Strong Law, the *Law of the Iterated Logarithm*, also seems to answer the other question of how large the interval is within which the frequencies are confined with increasing n: with probability 1 the frequency oscillates within an interval of length $([\ln \ln n]/2n)^{1/2}$, where ln signifies the natural logarithm, the inverse function to exp.

n tends to infinity, of $n^{-1}\Sigma X_i$, where the sum is from 1 to n, exists and is equal to p (as we pointed out in Chapter 2, its proof in this form requires the principle of countable additivity). Thus, the claim proceeds, the assumption that a sequence of coin tosses is approximately modelled as a sequence of Bernoulli trials (we shall return to this question in a moment) is sufficient to explain the convergence of the frequencies of heads and tails to within an arbitrarily small interval characteristic (because of the fixed value of p) of the experimental arrangement.

More than a pinch of salt is advised before swallowing this story. Even if we were to permit the use of countable additivity, and hence approve the Strong Law in its strongest form, it is not difficult to see that in itself it explains nothing at all, let alone sanctions the identification of probabilities with observed relative frequencies, since no empirical meaning has yet been given to the probability function P. Even were it true that a large value of P, *even the value 1*, attaches to a particular event cannot by itself explain any occurrence of that event: no statement of the pure probability calculus by itself makes any categorical prediction. Not only that: even were it the case that in some acceptable sense the Strong Law, or rather the modelling hypothesis implying it, that the coin tosses are an instantiation of a Bernoulli process, explained the convergence of relative frequencies, this by itself would be no ground for accepting the explanation offered. As we shall see in a subsequent chapter, there are infinitely many possible distinct and mutually inconsistent explanations of any observed effect. There has to be some independent reason for selecting any one of these rather than any other, which the Bayesian theory formally represents in terms of different prior probabilities.

We can also reject the frequently-made claim that the Strong Law shows why chance cannot be *defined* as limiting relative frequency—because the identity only occurs on a set of measure 1, and so it is possible in principle for the relative frequency not to converge, or to converge to some other value. But again, without any independent meaning given to P, or independent reason to accept the modelling hypothesis itself, this claim is empty. Some additional assumption linking statements of the probability calculus to physical reality is, on the contrary, *indispensable*. We have

made long-run frequency a definitional link, because that is the simplest procedure that avoids the objections. There have been other suggestions. One, which seems first to have been explicitly made by A.A. Cournot in the nineteenth century, and which often goes under the name of *Cournot's Principle*, is that sufficiently small probabilities should be assumed to be practical impossibilities (the same rule was also proposed by Kolmogorov in his 1950 monograph 1950). After all, the minute probability that the kinetic theory ascribes to ice forming spontaneously in a warm bath is typically taken as an affidavit that it won't happen in anyone's lifetime. Conversely, if it happened rather often, that would almost certainly be taken as a good indication that the kinetic theory is incorrect.

It should be clear, however, that without some qualification Cournot's Principle is false, for events with almost infinitesimal probabilities occur all the time without casting any suspicion upon the theories which assign them those probabilities: the exact configuration of air molecules in a room at any given time has a microscopic probability; so does any long sequence of outcomes of tosses of a fair coin (even with so small a number as twenty the probability of each possible sequence is already around one in a million). Can the Principle be qualified in such a way as to make it tenable? This is just what the well-known and widely-used theories of bivalent statistical tests of R.A. Fisher, and Neyman and Pearson (all believers in a long-run frequency account of statistical probabilities), attempt to do in their different ways. Unfortunately, as we shall show later, in Chapter 5, these attempts also fail.

But there remains the objection to a limiting frequency measure of objective probability that relative frequencies in finite samples deductively entail nothing whatever about the behaviour of limits, an objection reinforced by the fact that these limits only exist in an idealised modelling sense. The remarkable fact is that empirical data *do* nevertheless give us information about the limit, though to show how we need to add to that definition another feature of actual random sequences besides their apparent convergence to small neighbourhoods of some characteristic value. This is the fact that these sequences are *random*. The intuitive idea, which was made mathematically rigorous by the

American mathematician and logician Alonzo Church using the theory of computable functions which he had independently invented, is that a sequence is random if there is no algorithm, into which a knowledge of the first n members can be inputted, $n = 1, 2, 3, \ldots$, for discovering subsequences in which the probabilities (the limiting relative frequencies) differ: were such a possibility to exist then it could be exploited by a mathematically adept gambler to generate a sure profit.

According to Richard von Mises the two principles of convergence and randomnness, or immunity to gambling systems, determine a plausible mathematical model, which he called a *Kollektiv*, of what is empirically observed in the field of stochastic phenomena, either carefully cultivated in casinos, or occurring naturally in the passing on of genes from two parents, radioactive emissions, and so forth. One of their consequences is partitioning a *Kollektiv* into n-termed subsequences, for each $n = 1, 2, 3, \ldots$, generates a new *Kollektiv* of n-fold Bernoulli sequences, in which successive members are probabilistically independent, with the chance of any specified outcome occurring at the ith place the same for all i, $i = 1, 2, \ldots, n$ (von Mises 1964, pp. 27, 28). In other words, a *Kollektiv* with outcomes 0 and 1 (say) represents an infinite sequence of n-fold random samples in which the outcomes are characterised by the same chance, p, as in the original *Kollektiv*. This remarkable fact might seem to preclude the application of the theory to the important and extensive field of stochastic phenomena where independence fails, and which instead exhibit systematic probabilistic dependencies like Markov processes for example. This is fortunately quite untrue: all that follows is that the particular random variables X_i, $i = 1, \ldots, n$, defined on the set of all 2^n n-tuples $(x_1, \ldots, x_n)$, $x_i = 0$ or 1, where $X_i = x_i$, are independent. The random variables $Y_k = X_1 + \ldots + X_k$, $k = 1, \ldots, n$, are certainly not independent.

Later, we shall show how it follows from these properties of *Kollektiv*s that data from finite sequences can and do actually give information about the behaviour of limits. To do this we need first to develop the theory of epistemic probability, the *Bayesian* theory, for it is only within such a theory that this fact can be demonstrated. That it can be demonstrated at all is remarkable enough. That it needs a theory of epistemic probability to do so underlines the indispensability of the latter for providing a basis of inductive inference; that this basis is also a secure *logical* basis is even more remarkable. That is the programme for the remainder of this chapter.

3.b Measuring Uncertainty

To carry out this programme we shall first have to explain more exactly what epistemic probabilities are. The answers that have been given, at different times over the past three centuries, differ in the details, but one thing on which nearly everyone is agreed is that epistemic probabilities are numerical measures of uncertainty. So far, so good, except that it is not very far, and the question is where to go from there. Here opinions differ. The development we shall favour is the one that seems to us the most natural and definitely the simplest; it brings with it as a consequence that the rules of epistemic probability are nothing but rules of logic: not deductive logic, but a logic that is very closely kin to it. We shall show that the axioms of probability are *a logic of uncertain inference*.

The idea that there is an authentic *logic* of uncertain inference, complementing the deductive logic of 'certain' inference, has an ancient pedigree, extending back to the beginnings of the mathematical theory of probability in the seventeenth century. Leibniz in the *Nouveaux Essais* and elsewhere said so explicitly, and the idea runs like a thread, at times more visible, at times less, through the subsequent development of the epistemic view of probability, right up to the end of the twentieth century. Thus Ramsey: "The laws of probability are laws of consistency, an extension to partial beliefs of formal logic, the logic of consistency" (1931, p. 182). Ironically enough, however, Ramsey did more than anyone else to deflect it out of a logical path by choosing to embed his discussion not within the theory of logic as it was then being (very successfully) developed in continental Europe, but within the very different theoretical matrix of axiomatic utility, which Ramsey himself was the first to develop axiomatically, and of which more anon.

In this chapter we shall try to establish just this link between contemporary deductive logic and the laws of epistemic probability. We shall proceed in stages, the first of which will be to show

how it is possible to make numerical assessments of uncertainty, and then see what follows from the properties these have. *That* it is possible to make such assessments is hardly in doubt: people have been doing so for centuries, albeit indirectly, in terms of *odds*; for example:

> **SPEED**: Sir, Proteus, save you! Saw you my master?
> **PROTEUS**: But now he parted hence, to embark for Milan.
> **SPEED**: Twenty to one, then, he is shipp'd already. (William Shakespeare, *Two Gentlemen of Verona*.[3])

We say 'indirectly' in terms of odds because 'odds' traditionally means 'betting odds', and betting odds *directly* determine the ratio in which money changes hands in a bet, a fact which raises problems for any attempt to use the odds the agent gives or accepts as a measure of their uncertainty. The method suggested in de Finetti (1937, p. 102), of identifying your personal degree of uncertainty with the odds at which you would be prepared to take either side of a bet at arbitrary stakes will certainly not work, for reasons which have been well-documented and are by now very familiar (your willingness to bet will be sensitive to the size of the stake, to your attitude to risk, maybe to moral considerations, and to possible other 'external' factors), and which were historically a powerful factor in making people think that there was no alternative to the explicitly utility-based approach which we shall discuss in the next section. It is not possible to evade the problem by requiring that the stakes are always sufficiently small that they can be safely assumed to be approximately linear in utility, since the Dutch Book argument for the conditional probability axiom (Chapter 2, axiom (4)) requires that one be prepared to make a combined bet against b and on $a\&b$ with a stake on b which can in principle be arbitrarily large, and the Dutch Book argument furnishes the prudential reason for making your betting quotients

[3] We are grateful to Vittorio Girotto and Michel Gonzalez (2001) for drawing our attention to this quotation. A more recent example of the same thing is this: "The betting among physicists, however, was that there was an even chance that the SSC [Superconducting Supercollider] would find exotic particles beyond the Standard Model" (Kaku 1994, p. 183).

obey the probability axioms: if they don't, you can be made to lose come what may (for an elementary proof see Gillies 2000 pp. 59–64).

But citing odds does not necessarily indicate a corresponding propensity to bet that way, and Speed's 20:1 odds were probably not meant to be taken in any such way: they are more likely to be (for Shakespeare) his assessment of the relative chances of his master shipping versus not shipping (we are using the word 'chance' here in an entirely informal vernacular sense). Call such odds *chance-based odds*. It is important to keep in mind that these odds are conceptually distinct from betting odds: they are (so far at any rate) *judgments about the relative likelihoods* of the relevant proposition and its negation. That said, these odds are nevertheless *numerically* identical to a special type of such odds, the agent's *fair betting odds*, which is the reason that they are referred to as odds at all. What are fair betting odds? The answer, as old as probability itself, is that *betting at your personal, chance-based odds balances the risk (according to your own estimation) between the sides of the bet*. According to a standard definition (Mood and Graybill 1963, p. 276, for example), risk is expected loss, and expected loss is something that is straightforwardly computable, using the normalisations 20/21 and 1/21 of Speed's 20:1 chance-ratio as Speed's respective *personal probabilities* of his master having shipped and not having shipped: the word 'probabilities' is justified by the fact that these normalisations are numbers in the probability scale of the closed unit interval, summing to 1.

It is now simple to see that odds in this ratio do indeed balance the risk. Suppose Speed is observing two individuals actually betting on whether his master has shipped. One individual collects Q ducats from the other if the proposition is true, and loses R ducats to the other if it is false. It is easy to work out that, according to Speed's evaluation, the risk of this bet to the first individual is equal to the risk to the second if and only if $R{:}Q = 20{:}1$. Hence 20:1 is Speed's fair odds. Note that the bet is fair according to Speed *independently of the magnitude of* $Q + R$, *i.e. of the stake*: its fairness, or otherwise, depends only on the odds $R{:}Q$. So chance-based odds are also fair betting odds. Normalised betting odds are called *betting quotients*, and so normalised chance-based

odds are also *personal fair betting quotients*. Thus we have the equations:

> *Personal probabilities = normalised fair betting quotients = normalised chance-based odds*

There is one proviso in all this, applying when the truth-value of the proposition in question is not fully decidable. One of the major applications of the Bayesian formalism is to the problem of deriving posterior probabilities for scientific hypotheses (hence the title of this book!). Yet, as Popper became famous for pointing out, if they are sufficiently general these are at best only one-way decidable, being refutable (with the help of suitable auxiliary hypotheses) but not verifiable. Clearly, the only practical way to equalise risk in a bet on these is for it to be at zero odds. But these may not, and often will not, correspond to your chance-based odds. But all we have to do to re-establish equality is to make the proviso, possibly counterfactual, that any bet will be decided, if need be by an omniscient oracle (this is a fairly standard procedure and, since nothing in this account stipulates that the agent must or should bet at their fair odds, an unexceptionable one; cf Gaifman 1979, p. 134, n.4).

The condition of equal (and hence zero) risk is, of course, equal to that of equal (and hence zero) expected gain, a quantity that Laplace termed *advantage* (1820, p. 20); thus fair odds are those also that confer equal, meaning zero, advantage on each side of the bet. Appeal to zero expected gain as the criterion of fair exchanges has of course come in for a good deal of adverse comment, with the celebrated St Petersburg problem alleged to show the practice must either be abandoned on pain of inconsistency, or amended to a corresponding condition of zero expected *utility*. This piece of conventional wisdom is incorrect. True, the zero-expected-gain condition, plus the additivity of expectation, implies that any finite sum paid for the famous offer gives the advantage to the buyer (the offer is payment of $\$2^n$ if the first head in a sequence of indefinitely many tosses of a fair coin occurs at the nth toss, for $n = 1, 2, 3, \ldots$). But all that follows from anyone actually being willing to pay a substantial sum is that they are either foolish or they have money to burn.

What does not follow is that the bet is unfair, and to conclude otherwise is simply to conflate equity and preference. That these are conceptually distinct is evident from the fact that I may prefer, for a variety of reasons, to accept a bet at greater odds than those I believe are the fair ones—I might actually want my opponent to have the advantage, for example. Nevertheless the conflation is routinely made. Explicit in Savage's claim that "to ask which of two 'equal' betters has the advantage is to ask which of them has the preferable alternative" (1954, p. 63)[4], it has been a systematic feature of discussions of subjective probability throughout the twentieth century, undoubtedly assisting the acceptance of the orthodox view that defining a fair bet in terms of expected gain commits the error of assuming that money is linear in value. Of course it is not, but nothing in the zero-expected-gain account implies that it is, or indeed anything at all about the value that should be placed on a gamble (the point is made forcibly in Hellman 1997). Savage's claim is, moreover, as he half-confesses,[5] a very considerable distortion of the historical record. From the early eighteenth century, the zero-expected gain condition has been a legally enforceable condition of fairness in games of chance in lotteries and casinos, any divergence between the odds and the assumed chances being thought almost certain to be manifested in the average gains in repeated gambles failing to tend to equality. Condorcet pointed this out over two centuries ago by way of defending the fairness of the St Petersburg gamble (Todhunter 1886, p. 393), and in general it seems to have been for a long time regarded as a fact, certainly about repeated games of chance:

> We can actually 'see' the profits or losses of a persistent gamble. We naturally translate the total into average gain and thereby 'observe' the expectation even more readily than the probability . . . Certainly a gambler could notice that one strategy is in Galileo's words "more advantageous" than another. (Hacking 1975, p. 92)

[4] This is typical within the orthodox approach: "the fair price of a wager is the sum of money at which [the buyer] should be equally happy to have either a straight payment of the sum or the wager itself" (Joyce 1999, p. 13).

[5] "Perhaps I distort history somewhat in insisting that early problems were framed in terms of choice among bets" (1954, p. 63). Quite so. But every successful revolution tends to rewrite history in its own favour, and the utility revolution is no exception.

Be that as it may (the scare-quotes suggest less than complete confidence on the part of the author), the discussion shows one important sense of fairness of odds which *demonstrably* has nothing to do with any consideration of preference. We cannot, however, invoke presumptive tendencies in long sequences of independent repetitions, if for no other reason than that most of the propositions we will consider do not describe repeatable events at all: they are *theories* and *evidence-statements* whose truth, or falsity, are singular 'events' *par excellence*. Nor does the surrogate of *calibration* seem to offer an acceptable solution. Your probability assignments are calibrated if, for each probability value p, the proportion of true propositions in the class of all propositions to which you ascribe p is sufficiently close to p. To rule out unrepresentative swamping by many repetitions of the same event (like tossing the same suitably weighted coin) some condition of independence is required, but this and the other conditions needed on size and structure of the equiprobable reference classes beg more questions than are answered. Van Fraassen proves that calibration in some possible world ensures obedience to the probability axioms (1983), and Shimony proves a similar result in terms of 'reasonable' constraints on truth-frequency estimation (1993), but the assumptions required are so strong as to deprive the results of most of their significance.

We nevertheless believe that the old idea that fair betting quotients are identified with personal chance-based odds represents a sound enough basic intuition which, despite orthodox claims to the contrary, can be maintained consistently with acceptance of the phenomena of risk aversion, concavity of utility functions, and so forth. The one which begs fewest questions of all, and would be the one presented if our aim were unalloyed rigor, is R.T. Cox's (we shall describe it briefly later), but it also requires fairly sophisticated mathematics. At any rate, we shall continue with our much simpler development based on the traditional idea of fair odds being those corresponding to the agent's 'true' odds—traditional, that is, until becoming (unjustly) a casualty of the utility revolution. This is a fitting opportunity to pause and take a longer look at that revolution, whose tendentious redefinitions of fairness and equity we have considered, and rejected as question-begging. Even a rigorous axiomatic development is no protection against questions begged or, as we shall see in the next section, fundamental problems left unsolved.

3.c | Utilities and Probabilities

Inaugurated in Ramsey's seminal essay 'Truth and Probability' (1926), the utility-revolution was and still is widely regarded as culminating successfully thirty years later in Savage's classic work (1954). Savage analyses an individual's uncertainty with respect to any proposition a in terms of how he or she ranks their preferences for gambles involving a. Thus, if two gambles with the same payoffs but on different propositions are ranked differently it is assumed to be because the agent thinks one of the propositions more likely than the other (an axiom states that this preference must be independent of the *absolute* magnitudes of the payoffs). Given suitable constraints on the agent's preference ranking, this relation determines what Savage calls a 'qualitative probability' ordering on propositions which, within a sufficiently big space of possibilities (it must be infinite), is representable by a unique probability function.

Secured on what seemed like a rigorous axiomatic base, this utility-based account of epistemic probability by philosophers became so dominant after the publication of Savage's text that it is fair to call it now the *orthodox account*. It is certainly so among philosophers, promulgated in Earman's declaration that "degrees of belief and utilities have to be elicited in concert" (1992, pp. 43–44), and Maher's "You have subjective probability function p just in case there exists a utility function u such that your preferences maximise expected utility relative to p and u" (1990, p. 382). Yet dissent is, we believe, well-founded. To start with, the claim that a utility-based development is the only sure foundation for a theory of personal probability is not true: the path via utility theory is actually far from sure. There is, in particular, a profound problem with the way probabilities are supposed to emerge from preferences. The gambles relative to which the agent's qualitative probability ordering is defined are a subclass of the class of possible acts, and it is this more inclusive class on which the preference relation is defined. Formally, they are *functions* from states to consequences (including constant functions which for each consequence takes that consequence as value for all states), and a critical assumption is that the consequences are capable of being described in such a way that their value to the agent can be

regarded as constant across possible states. This constancy assumption is highly unrealistic,[6] and Savage himself acknowledges that in practical situations the value-relevant consequences of our actions, as we represent them to ourselves at any rate, will depend on states of affairs holding that as yet remain uncertain: we necessarily contemplate what he calls 'small worlds', whose states are coarsenings of the 'grand-world' states that determine the ultimate consequences of our acts. Knowledge of these being typically denied us, the consequences we actually envisage are therefore in reality 'grand-world' acts with consequences indeterminate in terms of the 'small world' states (1954, pp. 82–91). Since to every 'small world' act there corresponds a unique 'grand world' act, it is a natural consistency condition that the expected-utility ranking of the 'small world' acts is consistent with that of their 'grand world' counterparts. Yet, as Savage himself shows, there are 'small world'–'grand world' pairs in which the preference rankings are consistent with each other but in such a way that the 'small world' probability function is *inconsistent* with the 'grand world' probability (1954, pp. 89–90).

While in itself this might seem, according to taste, either merely mildly anomalous or something rather more serious, it nevertheless portends something definitely more to the latter end of the spectrum. For a development of Savage's example by Schervish *et al.* shows that the phenomenon of sameness of ranking with different probabilities can be reproduced entirely within one 'grand world' (1990 p. 845). Their result is not, as it seems to be, in conflict with Savage's well-known proof of the uniqueness of the probability-utility representation, because that proof assumes that the values attached to the consequences are fixed independently of the states. In the construction of Schervish *et al.* this is actually true for each probability-utility representation separately, but the values assigned consequences in one representation are state-dependent in the other, and vice versa. While the formal correctness of Savage's result is not impugned, therefore, the fact nevertheless remains that altering the way consequences

[6] Cf Aumann 2000, pp. 305–06. Savage's reply to him is hardly convincing (*ibid.* pp. 307–310).

are valued can, at any rate in Savage's account[7], actually alter the probabilities elicited from the same set of preferences. Jeffreys had earlier voiced scepticism about basing subjective probability on revealed preferences among uncertain options, precisely on the ground that such preferences are irreducibly "partly a matter of what we want, which is a separate problem from that of what it is reasonable to believe" (1961, p. 30), a judgment which now seems fully vindicated. Schervish *et al.* themselves conclude that the situation described above places "the meaning of much of this theory . . . in doubt" (1990, p. 846).

The path from utilities to determinate probabilities is, paradoxically, rendered even less firm by *improving* the conceptual foundations of Savage's theory. Jeffrey's decision theory (1964), published a decade after Savage's, replaces Savage's problematic state-action partition with a more appealing single algebra of propositions describing both the taking of actions and the possible consequences ensuing, generating thereby a modified expected-utility representation of preference ('desirability') whose probability-weights are conditional probabilities of outcomes given the action contemplated (in Savage's theory they are unconditional probabilities distributed over a partition of states). But a well-known feature of Jeffrey's theory is that the desirability-axioms permit a class of probability functions, and one so extensive that it contains functions which disagree in their ordering of the propositions.

It follows almost immediately that it is possible for an agent to order propositions by degree of belief consistently with the Jeffrey axioms *in a way not representable by any single probability function* (Joyce 1999, p. 136). True, the class of functions can in principle be narrowed down: the trouble is that all the attempts to do so cause disruption elsewhere or beg the question. For example, Bradley (1998) shows that uniqueness can be obtained by adjoining to the underlying algebra of propositions a class of conditionals satisfying the so-called Adams Principle (which says essentially that the probability of a conditional is a conditional probability). But then some famous results of Lewis (1976) imply

[7] And not just in that: Schervish *et al.* point out that the same problem afflicts Anscombe and Aumann's (1963) and de Finetti's (1974) theories.

that the resulting logic must be non-Boolean, to say nothing of these conditionals possessing very counterintuitive features: for example 'If Jones is guilty then Jones is guilty' turns out to have no truth-value if Jones is in fact not guilty. Jeffrey himself counselled adding further axioms for qualitative probability orderings (1974, pp. 77–78)[8], and it is well-known how to do this to secure uniqueness. But this strategy effectively sells the pass as far as the programme for determining subjective probabilities by appeal to properties of rational preference is concerned. The continued concentration of activity in the theory of rational preference should not conceal the fact that one of its historically principal objectives, the determination of personal probabilities by a suitable elicitation of preferences[9], is if anything farther from being achieved than ever.

Not only do they seem to provide no secure foundations for personal probability: decision theories like Savage's rather noticeably fail to discharge the function they themselves set, which is, typically, identifying criteria for making optimal choices among possible actions. Savage's theory is explicitly a theory of *rationality*, of "the behaviour of a 'rational' person with respect to decisions" (p. 7). The scare-quotes signify a degree of idealisation in the theory Savage is about to develop; to this end he calls it a *model*. But a model of what? It is increasingly appreciated that the theory's prescriptions are, at any rate in most interesting applications, beyond the capacity of even the most rational human agent to obey, a feature remarked in observations to the effect that the theory assumes an in-principle impossible "logical omniscience" (Earman 1992, p. 124), and that its characteristically sharp (point) probability-values are completely unrealistic. The objection is not substantially deflected by appeal to the idealising character of Savage's theory. A model of rational behaviour that makes no allowance for the fact that people are highly bounded reasoners, using irreducibly vague estimates of probability, is an inadequate

[8] A strategy endorsed by Joyce: "The bottom line is that Jeffrey-Bolker axioms need to be augmented, not with further constraints on rational preference, but with further constraints on rational belief" (1999, p. 137).

[9] It was the explicit goal of Ramsey, who proved, or more accurately sketched a proof of, the first representation theorem for utilities.

model; hence the various attempts over the last forty or so years to develop what Hacking called "slightly more realistic personal probability" (1967).

What these last objections reveal, we believe, is not that the standard formalism of personal probability is insufficiently human-centric, but that it is *misidentified* as a model, even a highly idealising one, of a rational individual's beliefs. On the other hand, if it is not that, what is it? It is suggestive to ask what other well-known theory has as its domain a full Boolean algebra of assertions (up to equivalence) and rules for distributing values over its elements. Answer: *deductive logic*. Nobody (presumably) would regard this as the theory of the mental workings of even an ideal human reasoner; it is a theory of necessary and sufficient conditions for inferences to be valid and sets of sentences to be consistent, whose connection with *human* reasoning exists only in the occasional ability to map in a more or less faithful way inferences actually considered into the set defined and evaluated within the model. The formalism of epistemic probability suggests, we believe, a similar conclusion: that formalism is a model of what makes a valid probabilistic inference, not of what ideally rational agents think or ought to think in conditions of uncertainty. Of course, so far the kinship between epistemic probability and logic is no more than suggestion, and certainly does not amount to a proof, or even a semi-rigorous argument, to that effect. The differences between probability-values, inhabiting as they do an interval of real numbers, and the two truth-values 'true' and 'false', as well as their difference of interpretation, shows that any proof of identity between the two models is out of the question. But acknowledging their distinctively probabilistic features does not mean that in some relevant sense the laws of probability are not sufficiently close to those of deductive logic to merit being assigned the status of logic.

Indeed, we believe such a logical interpretation is not only possible (and the rest of this chapter will be devoted to arguing the case for this), but that *only* such an interpretation is capable of accounting in a natural way for those features of the formalism of epistemic probability which are otherwise highly anomalous. For example, on the logical view the charge of 'logical omniscience' is clearly misplaced, for there is no knowing or reasoning subject,

even an ideal one, appealed to or legislated for. The use of 'sharp' probability values is also easily justified as a simplifying move analogous to the adoption of 'sharp'—that is, strictly two-valued—truth-values in the standard deductive models. These latter would also be judged unrealistic were their function that of mirroring accurately the semantics of natural languages where, as the many varieties of Sorites demonstrate, truth is typically vague if not actually fuzzy (nor is even mathematics, that alleged paradigm of precision, immune, a fact most notably pointed out by Lakatos, who used the vagueness in the concept of a polyhedron as the starting point from which he began a celebrated investigation into the foundations of mathematics (1963); and opinion is still far from unanimous on the meanings of 'set', 'function', 'continuous', 'number', and other basic mathematical notions). Their role is not that of mirroring natural-language semantics, however, but of playing a central role within simplifying models of valid deductive inference whose payoff is *information*, about the scope and limits of deductive systems: witness the significance given the classic theorems of Gödel, Löwenheim, Skolem, Church, Tarski, Cohen, and others. Similarly, sharp probability values are justified by the explanatory and informational dividends obtained from their use within simplifying models of *uncertain* inference. And as we shall see in the remainder of this book, these are very considerable.

One—indeed the principal—question we must answer is why the probability axioms should be regarded as constraints on the distribution of fair betting quotients. One well-known way of trying to justify the probability axioms within a general betting-odds approach appeals to a purely arithmetical result known as the Dutch Book Theorem. First proved by de Finetti (1937), its content is this: *a function P whose values are betting quotients cannot generate, for any set of stakes, a positive net loss or gain independently of the truth-values of the propositions bet on, if and only P satisfies the finitely additive probability axioms*. A system of betting quotients invulnerable to such a forced loss de Finetti termed 'cohérent', usually translated into English as 'coherent' though it is also translatable as 'consistent'. Thus the theorem states that coherence for belief-functions measured in this way is

equivalent to their satisfying the probability calculus. For a simple proof of the theorem see Gillies 2000, pp. 59–64.

As mathematics, the result is unquestionable. What is problematic about it is why being a guarantee of invulnerability to a forced loss should authorise the probability axioms to constrain the distribution of personal fair betting quotients whose definition, as we have taken pains to stress, implies no propensity of the agent whose beliefs they characterise to bet at the corresponding odds, or even to bet at all. There is an answer to this question, as we shall see later in section **e**, but at the present stage of the discussion it is not obvious. In what follows we shall try a different approach.

3.d Consistency

Ramsey and de Finetti both regarded the probability axioms as consistency constraints. However, de Finetti identified probabilistic consistency with coherence (1937, p. 102), which rather begs the question, while Ramsey saw consistency as a property of sets of preferences. That way, of course, is the orthodox way of utility theory, which we have repudiated. Nevertheless, despite this unpromising start, we shall persist with the claim that the probability axioms are laws of consistency. Indeed, we shall argue that they are laws of consistency in very much the same way that the laws of deductive logic are laws of deductive consistency.

Now this might seem an even more unpromising proposal than either de Finetti's or Ramsey's, since to start with deductive consistency is a property of sets of sentences, not of assignments of numerical fair betting quotients to propositions. The proposition-sentence difference is not the problem here, since for our purposes they can be regarded as the same things. The problem, or what seems to be a problem, is that in one case consistency is predicated of *assignments* of number-values, and in the other of *sets* of sentences or propositions. Perhaps surprisingly, this gulf is less formidable than it looks. At any rate, it is easily bridged by recognising that the apparently distinct notions are merely subspecies of a single more general concept, the familiar mathematical

concept of consistency defined as the *solvability of sets of equations*. To see why this is so, firstly note that a set of assignments of number-values to a Boolean function is in effect a set of equations. Thus, considering a *consistent* set K of assignments to a belief function Bel( ) and a conditional belief function Bel(|), Paris writes

> Consistent means that, with whatever additional conditions currently apply to Bel( ), Bel(|), there is a solution satisfying these and the equations in K. (1994, p. 6)[10]

The next step consists in observing that deductive consistency is also really nothing but solvability in this sense. This is easiest to see in the context of a popular deductive system for first order, and in particular propositional logic, the *semantic tableau* or *tree* system.

This is specifically designed to be an efficient test for the deductive consistency of sets of sentences. Here is a simple example: we want to test the set $\{a, a \rightarrow b, \sim b\}$ for consistency. The test consists of writing these sentences so

$$a$$
$$a \rightarrow b$$
$$\sim b$$

and beneath $\sim b$ appending the tableau rule for the unnegated conditional $a \rightarrow b$

$$/ \quad \backslash$$
$$\sim a \quad b$$

We have two branches from the root, on both of which are a sentence and its negation. The tree is now closed, signifying that the set is inconsistent, and the test is complete (the classic text is Smullyan 1968; more elementary texts are Jeffrey 1989 and Howson 1997).

[10] Paris takes K more generally to be a set of linear constraints on belief functions.

Essentially the same tableau, or tree, could however have been written in the 'signed' form

$$v(a) = 1 \quad \text{(i)}$$
$$v(a \rightarrow b) = 1 \quad \text{(ii)}$$
$$v(b) = 0 \quad \text{(iii)}$$
$$/ \quad \backslash$$
$$v(a) = 0 \quad v(b) = 1$$

where 1 signifies truth and 0 falsity, and v is a truth-valuation function[11]. What we see here is most revealing: the signed tree shows that the initial assignment represented by the equations (i)–(iii) is *overdetermined*. Those equations are unsolvable over the set of propositional variables: there is no assignment of values to a and b which satisfies them. Conversely, had the initial set been consistent, a corresponding complete and *open* (signed) tree could be constructed which would exhibit at least one single-valued extension of the initial valuation to all the sentences of the propositional language (Howson 1997, pp. 57–60). We should note that this phenomenon is not restricted to propositional logic: it is straightforward to show that signed trees perform an analogous role for full first order logic.

We see, then, that the signed tree method shows explicitly *that the consistency or inconsistency of a set of sentences is equivalent to the satisfiability or unsatisfiability of sets of equations*. It should also be evident that all signed tableaux can be converted into unsigned ones (by first removing all the truth-values and then negating any sentence to which 0 had been assigned), and conversely. In other words, we see that the ordinary concept of deductive consistency of a set of sentences in some interpretative structure is equivalent to the solvability of an assignment of truth-values.

We can now paraphrase Paris's definition of a consistent set K of assignments appropriately for this deductive example: K (here the set {(i), (ii), (iii)}) is consistent if, with whatever additional conditions apply to $v( )$, there is a solution satisfying these

[11] Smullyan 1968, to whom the definition of signed tableaux is due, signs them with T and F; we equate the sentences to 1 and 0 to emphasise the equational character of the tableau.

conditions and the equations in K. It follows that truth is involved in the concept of deductive consistency *only via the constraints*, and suggests that corresponding to different types of constraint on 'values' attached to propositions we shall obtain different types of logic: so for example the logic of consistent assignments of *truth-values* subject to the constraints given by a classical truth-definition is deductive logic; the logic of consistent assignments of *fair betting quotients* determined by appropriate constraints on them we can legitimately call *probability-logic*.

The Big Question, of course, is what these 'appropriate constraints' are. We already have one: the scale of probabilities, since we are dealing with normalised odds, which occupy the closed unit interval (and add over pairs a, $\sim a$). And we also have the primitive idea that those numbers represent fair betting quotients, reflecting assessments of the agent's 'true' odds. The immediate question is where to go from there.

A clue is suggested by the fact that since we are discussing consistent *distributions* of probabilities, we should presumably be looking for some collective condition or conditions. But what collective condition? Joyce has recently argued that the probability axioms follow from the condition that an overall assignment is 'truth-directed', but the conclusion depends on representing 'true' by 1 and 'false' by 0 (1998): reverse this entirely conventional proxying, and the result fails. But let us try to exploit the idea of fairness a bit more. A very plausible condition is that fair gambles should not become collectively unfair on collection into a joint gamble, and it also points in the desired direction since (a)–(c) below show that the laws of probability pretty well all follow from the condition that a finite set of fair bets is fair (de Finetti first showed this in his 1972, p. 77). A problem is that if the criterion of fairness is zero-expected gain, we have as yet no information on how to aggregate expectations on sums of bets; that they add over finite sums of random variables is a consequence of the finitely additive probability axioms, but to invoke the axioms at this stage would clearly beg the question.

However, there is nothing to stop us adding as an independent assumption that the class of fair bets, bets whose betting quotient is the agent's fair betting quotient, is closed under finite sums. De Finetti himself does so, in the form of a *definition* of fairness for finite sums of bets (*ibid.*) For him, however, this 'definition' is in effect a consequence of insisting monetary stakes be kept small:

in the limit of small sums utility-gains are equal to monetary ones, and utility functions compensate for the departures from additivity (which he calls 'rigidity' [1974, p. 77]) caused by risk-aversion where the stakes are in monetary units. But that, of course, is a utility-based justification. Our own, where no condition is placed on the size of the stakes, is merely that it is a natural assumption that chance-based odds do not 'interfere' when combined in finite sums of bets, one arguably as analytic in its way as the deductive assumption that a conjunction is true just in case all its conjunctions are. We do not even need the principle in its generality, only the following special case:

(C) If the sum of finitely many (and in fact we never need to exceed two) fair bets uniquely determines odds O in a bet on proposition a, then O are the fair odds on a.

(C stands for 'Closure', that is, the fact that it is a closure principle on fair bets; "uniquely" means 'independently of choice of stakes'; on those occasions when we will use (C) it is not difficult to see that the condition is satisfied.)

3.e The Axioms

To keep things simple we shall now follow de Finetti 1972 in identifying a proposition a with its indicator function: i.e. the function taking the value 1 on those states of affairs making a true and 0 on those making a false. This means that from now on we are dealing with an *algebra* (of indicator functions). The *random quantity* (de Finetti's terminology) $S(a - p)$ then represents an unconditional bet on/against a with stake S (S positive for 'on' and negative for 'against') and betting quotient p, paying $S(1 - p)$ when a is true and $-pS$ when a is false. A bet on a conditional on the truth of another proposition b is a bet on a if b is true, called off if b is false. If the stake is S it is therefore the quantity $bS(a - p)$. p is your *conditional fair betting quotient* if it is your fair betting quotient in a conditional bet.

The following facts are easy to establish:

(a) If p is the fair betting quotient on a then $1-p$ is the fair betting quotient on $\sim a$. This anyway we know to be the

case, since as we saw earlier p and $1 - p$ are the normalisations of the agent's chance-based odds.

(b) If a and b are disjoint with fair betting quotients p and q respectively then $p + q$ is the fair betting quotient on $a \vee b$.

(c) If p and q are the fair betting quotients on $a\&b$ and b respectively, and $q > 0$, then p/q is the conditional fair betting quotient on a given b.

Proof:

(a) $S(a - p) = -S(\sim a - [1 - p])$. Now use (C).

(b) $S(a - p) + S(b - q) = S(a \vee b - (p + q))$. Now use (C).

(c) If both $p,q > 0$ then there are nonzero numbers S, T, W such that $S(a\&b - p) + T(b - q) = aW(a - p/q)$ (T/S must be equal to p/q). Now use (C). The restriction to $p > 0$ can be eliminated by noting that if $p = 0$ then $S(a\&b - p) = Sa\&b = Sab = bS(a - 0)$.

A point of interest is that (C) can be seen as a more general additivity principle: (b) tells us that it underwrites the familiar additivity principle for probabilities, but (c) tells us that it also underwrites the quotient 'definition' of conditional probabilities. This kinship between two of the fundamental laws of probability usually regarded as being distinct from each other is of interest whatever view one might take of the nature of probability itself.

We now come to the main result of this chapter. Let $\mathfrak{R}$ be some algebra of propositions, assumed fixed for the following discussion (as we shall see shortly, the result does not depend on the particular choice of $\mathfrak{R}$), and let K be an assignment of personal fair betting quotients (including conditional betting quotients) to a finite subset $\mathfrak{R}_0$ of $\mathfrak{R}$.

Theorem: K can be extended to a single-valued function F on $\mathfrak{R}$, where F satisfies (C), if and only if K is the restriction of a finitely additive probability function on $\mathfrak{R}$.

Proof.

We shall actually prove something slightly stronger, namely that the (nonempty) class of all single-valued extensions of K to $\mathfrak{R}$ which satisfy (C) is exactly the (nonempty) class of finitely additive probability functions on $\mathfrak{R}$ which agree with K on $\mathfrak{R}_0$.

(i) 'Only if'. It follows almost immediately from (a)-(c) that any extension of K satisfying the conditions stated is a finitely additive probability function on $\mathfrak{R}$ (the additivity principle is obvious, and we also have that the logical truth t is equal to $a \vee \sim a$ for any a, and $P(a \vee \sim a) = P(a) + P(\sim a) = P(a) + 1 - P(a) = 1$; the corresponding condition that $P(\perp) = P(a\&\sim a) = 0$ follows immediately).

(ii) 'If'. Suppose that K is the restriction of some probability function P on $\mathfrak{R}$. It is sufficient to show that if $p_1, \ldots, p_k$ are the values given by P to $a_1, \ldots, a_k$, and a sum of bets $X_1, \ldots, X_k$ at those betting rates is a bet on some proposition a with betting quotient p, then $p = P(a)$. For then P itself will be such a single-valued extension of K to $\mathfrak{R}$ satisfying (C). Suppose the antecedent is true. Since the expectations of the X_i with respect to P are all 0, it follows that the expected value of $Z = \Sigma X_i$ is 0 also, by the linearity of expectations. Hence if $Z = S(a - p)$ for some proposition a then the expected value of the right hand side must be zero, in which case $p = P(a)$ (it is straightforward to show that the result continues to hold if one or more of the bets is a conditional bet). QED

Corollary 1 K is consistent just in case K is the restriction of a finitely additive probability function.

We can now push the analogy with propositional logic even farther. Define a *model* of a consistent assignment K to be any single-valued extension of K to $\mathfrak{R}$ which satisfies (C), and say that an assignment K' is a *consequence* of another assignment K just in case every model of K is a model of K'. We then have:

Corollary 2 K' is a consequence of K just in case every probability function which extends K assigns values to members of $\mathfrak{R}$ as specified by K'.

Discussion. To say that P is a probability function over $\mathfrak{R}$ is to say that the pair $(\mathfrak{R},P)$ is a system implicitly defined by the familiar axioms of mathematical probability: the probability calculus is just their deductive closure. Indeed, having now situated the enterprise of uncertain reasoning firmly within the province of logic, we can construe those axioms as logical axioms, mediating the drawing of consequences from specific sets of premises, which we can regard as assignments like K (rather like in the signed semantic tree representation of deductive inference, where the initial sentences are also value-assignments). A further corollary is that valid uncertain reasoning, construed as reasoning involving personal probabilities, is formally probabilistic reasoning—reasoning in accordance with the rules of the probability calculus. Indeed, suppose that we can construct a representation of $\mathfrak{R}$ in a deductive system, in which we can also express the ideas of a probability system, and of an arbitrary assignment of real numbers to members of $\mathfrak{R}$. Such a deductive system is a *probability calculus* based on $\mathfrak{R}$ whose axioms, since they serve only to mediate a consequence relation, can accordingly be regarded as having logical status: they are in effect logical axioms in a theory of valid probabilistic consequence.

And we can say more. The only 'facts' about the general properties of a probability function on an algebra are the logical consequences of a suitable axiomatisation, together with a suitable representation of the algebra. In a first-order framework these logical consequences are just the formally provable consequences, in which case the theorem above can be seen as a *soundness and completeness theorem for probabilistic consequence*: it implies that *K' is a consequence of K just in case there is a proof of K' from K in the corresponding formal calculus*. A concrete example is provided by Paris (1994, pp. 83–85), where the domain of the agent's uncertainty is the set SL of sentences in the language L generated by a finite set of propositional variables under the connectives $\&$, v, and $\sim$. Paris presents a formal deductive system Π for sequents of the form $K \mid K'$ (to be read "*K entails K'*"), where

K and K' are now finite sets of linear constraints (i.e. sets of the form

$$\sum_{j=1}^{k} u_{ij}B(a_j) = v_i, i = 1, \ldots, m$$

where $B(a_j)$ is the agent's probability assignment to a_j, and where the axioms and rules of inference are explicitly specified. Paris then proves the following theorem:

For any pair K, K', the set of all probability functions extending K to SL is contained in those extending K' if and only if the sequent $K \mid K'$ is a theorem of Π.

(We can straightforwardly adapt the result to a corresponding algebra of propositions, or their indicator functions, by the usual device of identifying logically equivalent members of SL.) Since the antecedent of Paris's theorem is equivalent to the condition for probabilistic consequence as defined above, that theorem is indeed such a soundness and completeness theorem relative to an explicitly-defined syntactical structure.

In the light of this discussion there seems no longer any room for doubt that the probability calculus is interpretable as, and we would argue should be interpreted as, a set of consistency constraints on distributions of personal probability in a way strongly analogous to the way that standard deductive rules are consistency constraints on truth-value assignments.

Corollary 3 K is consistent if and only if no system of bets with odds given by K will lead to certain loss.

This follows immediately from the result above and a sharpening of the Dutch Book theorem by de Finetti (this states that a non-negative real-valued function F on $\mathfrak{R}_0$ can be extended to a finitely additive probability function on any algebra $\mathfrak{R}'$ including $\mathfrak{R}_0$ iff the betting quotients obtained from F are coherent on $\mathfrak{R}_0$ (de Finetti 1972, p. 78; the proof is by induction on the well-ordered set obtained from $\mathfrak{R}_0$ by adding an arbitrary proposition at each stage; if $\mathfrak{R}'$ is uncountable the proof uses transfinite induction on

a well-ordering of $\mathfrak{R}'$, for the existence of which the Axiom of Choice has to be assumed).

Discussion. While coherence is now not a primary, or in itself any, justification of the probability axioms, merely a corollary of the consistency of the assignments of fair betting quotients, this result is of considerable importance from another point of view, namely *sanctions*. It is precisely the issue of sanctions for violation of the rules of the probability calculus that might be thought to present a big *disanalogy* between probabilistic and deductive logic. Disobeying deductive rules of consistency invites the possibility of accepting a set of statements as simultaneously true which in fact cannot be. This is, of course, a feature peculiar to deductive logic, though it is one whose importance is, we believe, greatly exaggerated. The constraints usually regarded as determining deductive consistency are the rules of a classical truth-definition. Yet in ordinary discourse these rules are frequently and, depending on the case, systematically violated—fortunately: otherwise we could immediately infer the existence of God from the apparently true premise that it is not true that if He exists we are free to do as we like. That the negation of a material conditional implies the truth of its antecedent is an artefact of its definition as a bivalent truth-function, resulting in classical first-order logic providing a notoriously poor model of conditional reasoning (this is a fact that is not emphasised in the standard textbooks, including one by one of the authors, Howson 1997). And that is merely an extreme case: the commutativity of conjunction is frequently 'infringed' without people declaring the speaker inconsistent: for example, 'she got married and had a baby' does not necessarily mean the same as 'she had a baby and got married'. These facts, together with the development and serious candidacy of certain types of non-classical deductive logic, suggest that it is merely naïve to believe that deleterious consequences are any more likely to flow from accepting a classically inconsistent set of sentences. For that matter, none of our current foundational scientific theories, including that for mathematics itself, are or even can be known in any non-question-begging way to be classically consistent: Gödel's famous Second Incompleteness Theorem states that there is no consistency proof for them which does not require

stronger premises than they themselves possess—unless they are inconsistent.

So, surprisingly for those who accept the conventional wisdom, the case that failure to the canons of deductive consistency places the violator in any sort of jeopardy has yet to be made, and in view of the preceding observations it is most unlikely that it ever will be. By contrast (and turning the conventional wisdom on its head), the corollary above shows that there are much more palpable sanctions against violating the rules of probabilistic consistency. Personal probabilities determine fair betting quotients, and being inconsistent in the use of these means potentially inviting certain loss. It is no objection to say that the potential for certain loss is merely that, a potentiality: also 'merely potential' are the deleterious consequences supposedly arising from accepting an inconsistent set of sentences.

> *Corollary 4* If K is consistent then it is so independently of the particular $\mathfrak{R}$ chosen; that is, K has a single-valued extension satisfying (C) to *any* algebra which includes $\mathfrak{R}_0$.

This follows immediately from the theorem of de Finetti stated in Corollary 4. Note that it is certainly not the case if one substitutes 'countably' for 'finitely', since it is well-known that not every subset of the unit interval, for example, is measurable (assuming the Axiom of Choice).

Discussion. There is an important analogous property of deductive consistency: the consistency of a truth-value assignment depends only on its domain; equivalently, a set of sentences is deductively consistent, or not, independently of the constitution of the language in which it is included. This *local* character of deductive consistency, and by implication deductively valid inference, is therefore mirrored in the account of probabilistic consistency (and probabilistic consequence; see the following corollary). The importance of focusing on local assignments to sets of propositions which are not necessarily, or even usually, closed algebras, is always stressed by de Finetti in his writings, and it is no accident that he often chose a logical or quasi-logical vocabulary to describe his results: he certainly saw what he was doing as a part of a more general theory of logic. It is no accident either that

characteristic properties of deductive logic carry over to its probabilistic analogue, and the fact that in Carnap's early systems of inductive logic the 'degree of partial entailment' between h and e *did* depend on the structure of the language of which h and e were a part is a significant factor against them.

Historical Note. Most people who have tried to construct logical theories of probability regard them as assigning probabilities to formulas in a formal language. Interesting results have come out of this research. For example, Gaifman proved a 'logical' analogue of the Extension Theorem for measures: he showed that any probability function defined on the quantifier-free sentences of a first order language L with denumerably many constants has a unique extension to the full set of sentences of L, which satisfies the condition that the probability of an existentially quantified sentence is the supremum (in the corresponding Lindenbaum algebra) of the probabilities of its instantiations with constants (Gaifman 1964; this condition is now known as the Gaifman condition). Similar results were proved for infinitary languages by Scott and Krauss (1966), and 'logical' analogues of Bayesian convergence-of-opinion theorems have been obtained by Gaifman and Snir (1982).

Our view is that the assignment of probabilities to formal sentences or more generally formulas is neither a necessary nor a sufficient condition for an authentically logical interpretation of probability. It is not sufficient, because the case for making epistemic probability authentically logical arguably requires finding some way of interpreting the rules of assignment, the probability axioms, as recognisably *logical* rules, and this is not done simply by assigning non-negative real numbers to formulas in a way that satisfies the probability axioms. Nor is it necessary, we believe, because the account given above provides an authentically logical interpretation of the probability calculus. It is actually quite compatible with the numbers-assigned-to-formal-sentences one, since we can if we wish take the algebra of propositions to be the sets of realisations (models) of the corresponding sentences (the algebra is then isomorphic to the Lindenbaum algebra of the language), but it is much more general in that it includes in its scope propositional structures not explicitly linguistic in character.

Standard formal languages are actually rather restrictive when it comes to describing mathematical structures. First-order languages are well-known to be deficient in this respect, being unable to characterise up to isomorphism even very simple structures like the natural numbers, or indeed any infinite structure at all, so the increase in determinateness provided by higher-order languages has to be traded against the loss of a complete axiomatisation of the underlying logic.

The logical perspective provides compelling answers to some of the most frequently raised problems, or what are taken to be problems, in the literature on Bayesian probability. One of the most prominent of these is the so-called 'problem of logical omniscience'. This is that the incorporation of deductive conditions in the probability axioms—for example the condition that all logical truths have probability one, and all logical falsehoods have probability zero—means that the Bayesian agent has to be 'logically omniscient' to apply those axioms correctly. This is a serious problem if the Bayesian theory is supposed to be a model of rational agents, but no problem at all, as we noted earlier, if the formalism of Bayesian probability is seen merely as a model not of reasoners but of valid reasoning in itself, in which the axioms are consistency constraints in much the same way that the axioms of first order logic are consistency constraints, or in other words as constraints on the consistent distribution of certain quantities independently of any agent's thought-process or cognitive abilities.

These observations also effectively dispose of another frequently-made objection, that the Bayesian theory requires that all the possible hypotheses, and all the possible evidence statements, one might ever consider must all be explicitly present in any calculation. The theory was invented by working scientists (we have mentioned Bayes, Laplace, Poincaré, Jeffreys, Jaynes, Cox, Savage, Lindley and others) to help them codify, understand and justify the principles they appeared to be applying in making inferences from data, and it seems hardly likely that they would have produced a theory impossible in principle to use. At any given time there may be a limited number of hypotheses in play that one would want to consider seriously, or find relevant to the problem at hand. Probability is distributed over these, with perhaps some kept in hand to be assigned to 'other causes' when and

if these ever come to be considered (this is the view propounded by Abner Shimony under the name of "tempered personalism" (Shimony 1993, pp. 205–07). It may be that new information, or merely more recent consideration, causes one to want to redistribute these probabilities, and not just by conditionalisation (that is, after considering new data); perhaps the information is logical, and one sees logical dependencies where before one failed to; perhaps new hypotheses are invented, or discovered, but then again perhaps not, and one just feels dissatisfied that one's original assignments reflect what reflection deems they should. Well and good; nothing in the Bayesian theory says that this should not be allowed; it would be absurd to insist on any such statute of limitation, and the theory does not do so. Its propositions are there to be consulted, but in a sensible way, understanding that, just like those of deductive logic, they are a servant and not a master.

3.f The Principal Principle

Relative to the constraints that fair betting quotients lie in the closed unit interval, together with the collective condition (C), the probability axioms are necessary and sufficient conditions for assignments of personal probability to be consistent. But there is one aspect in which the constraints themselves are incomplete. Suppose that a describes one of the possible outcomes of an observation, over which there is a well-defined objective probability distribution P^*. Our assumptions (a) that objective probabilities are numerical measures of tendencies scaled in the closed unit interval, and (b) that personal probabilities reflect personal assessments of relative likelihoods, strongly suggest that conditional on your sole information that the data source possesses a definite tendency to produce outcomes as measured by P^*, your personal probability of an a-type outcome should be equal to $P^*(a)$. We can express this principle formally as follows. Let a be a generic event-descriptor, and a_0 be the prediction that a specific instance of a will be observed when suitable conditions have been instantiated. Then

$$P(a_0 \mid P^*(a) = r) = r \qquad (2)$$

(2) is a version of the principle traditionally known (in the English-speaking world) as the Principle of Direct Probability, whimsically redubbed *The Principal Principle* by David Lewis in a well-known paper (1980).

It is (2) that enables posterior distributions to be calculated, via Bayes's Theorem, over values of these chances: without (2), the subjective probability formalism would be empty of any methodological content. Indeed, one of the most striking results obtained by using it provides the solution to a problem raised earlier, with the promise that it would be solved. This is the problem of how defining objective probabilities as limiting relative frequencies allows sample data, or records of what happens in finite sequences of observations, to provide any information at all about objective probabilities. After all, not only is any behaviour in an initial segment of an infinite sequence compatible with any limiting behaviour, but we can be assured that infinite *Kollektiv*s do not even exist, given that the lifetime of the Universe, at any rate of that bit of it in which a coin, say, is repeatedly tossed, is presumably finite. It seems frankly almost beyond belief that such information could be provided. *But it can, as we shall now see.*

There are three steps in the explanation of how. The first is to recall that, because they are defined in terms of limiting relative frequencies, objective probabilities satisfy the axioms of the probability calculus. The second is also to recall that *Kollektiv*s are Bernoulli sequences: the probability of any outcome at any point is independent of the outcomes at other points, and is constant from point to point. The third and final step is to use the resources of our theory of epistemic probability. To this end, let h state that the X_i are generated by a chance mechanism, in other wsords, for some value of the chance that $X_i = 1$, and h_p state that that chance is p. Finally, let $e(r; n)$ be the statement that r ones are observed in a sample of size n. Suppose that we have a fairly smooth prior density function $h(p)$ for p, whose integral of $h(p)$ between 0 and 1 is $P(h)$, since h just says that h_p is true for some value of p. By Bayes's Theorem and the Principal Principle the posterior density $h(p \mid e(r,n))$ is proportional to ${}^nC_r\, p^r(1-p)^{n-r}h(p)$. As n grows large this density becomes dominated by the likelihood $p^r(1-p)^{n-r}$ (for large n the logarithm of the likelihood is approximately a constant

times n), which is a maximum at $p = r/n$ and falls to zero very rapidly away from the maximum, leaving the posterior density of points close to the relative frequency proportionally increased more as the sample grows large. Another way of putting this is to note that the approximate likelihood ratio $(r/n)^r(1 - (r/n))^{n-r} + p^r(1-p)^{n-r}$ increases without bound for p outside a small interval which tends to 0 as n grows large, and if one accepts the plausible *Likelihood Principle* (see Chapter 5, p. 156), which says that the information gained from the sample is expressed in the likelihood function, then the relative frequency in finite samples is certainly highly informative about the value of p. Either way, we have demonstrated, within the Bayesian theory of epistemic probability, that finite sample data do provide information about a chance parameter *even when this parameter refers to an in-principle unobservable limit*.

The posterior probability of any interval of values for p will also depend on the prior probability $P(h)$ of h, since the prior density $h(p)$ must integrate to $P(h)$ between the limits $p = 0$ and $p = 1$. $P(h)$, recall, is the prior probability that the data source will generate a *Kollektiv*, and so it is the prior probability of a one-dimensional *statistical model*, with undetermined parameter p. Von Mises was extremely careful to base his theory on what he took to be two already established empirical laws about random phenomena: (i) that they are truly random, in the sense of being immune to gambling systems, and (ii) that their relative frequencies exhibit convergent behaviour. *In Bayesian terms, von Mises (though he was not, at any rate consciously, a Bayesian) was in effect arguing for a considerable prior probability to be attached to the modelling hypothesis h*. The posterior *probability* of h is of course obtained by integrating the posterior density between the limits 0 and 1.

In the following chapters we shall apply the Bayesian tools to more sophisticated, and also practically more important, statistical hypotheses than toy examples like h, involving models with arbitrary numbers of adjustable parameters. But in terms of establishing one of the central claims of this book, that an appropriate theory of epistemic probablity permits a demonstration of the fact that finite sample data can indeed provide information about infinite limits, the example above is anything but trivial.

3.g Bayesian Probability and Inductive Scepticism

David Hume produced a celebrated circularity argument, that any justification for believing 'that the future will resemble the past' in any specified manner must explicitly or implictly assume what it sets out to prove. We believe (the claim is argued at length in Howson 2000) that Hume's argument shows informally and in general terms that a valid inductive inference must possess, in addition to whatever observational or experimental data is specified, at least one independent assumption (an inductive assumption) that in effect weights some of the possibilities consistent with the evidence more than others. The Bayesian theory viewed in a logical perspective endorses this view in a quite satisfactory way, and in so doing reveals a further illuminating parallel with deductive logic. A nontrivial conclusion (one which is not itself a theorem of logic) of a deductively valid inference depends on at least one nontrivial premise. Similarly, a nontrivial conclusion (one which is not a theorem of probability) of a valid probabilistic argument depends on one or more nontrivial probabilistic premises. And just as the logical axioms in Hilbert-style axiomatisations of classical logic are regarded as empty of factual content because they are universally valid, so is the same true of the probability axioms in the view we have been advocating. They too are logical axioms, empty of factual content because universally valid.

Putting all this together, we can derive a probabilistic analogue of the celebrated conservation result of deductive logic, that valid deductive inference does not beget new factual content, but merely transforms or diminishes the content already existing in the premises. So too here: valid probabilistic inference does not beget new content, merely transforming or diminishing it in the passage from premises to conclusion. This was well understood by de Finetti:

> *The calculus of probability can say absolutely nothing about reality* . . . As with the logic of certainty, the logic of the probable adds nothing of its own: it merely helps one to see the implications contained in what has gone before. (1974, p. 215; emphasis in the original)

We can say more. The 'synthetic' premises in a probabilistic inference are generally prior, or unconditional, probabilities, and because their exogenous nature is explicitly acknowledged within the so-called subjective Bayesian theory they are often seen as its Achilles heel. Hume's argument enables us to view them in a less unfavourable light, for it implies that some degree of indeterminacy is a natural and indeed inevitable feature in any adequate theory of valid inductive inference. Far, therefore, from being a disabling feature of the subjective Bayesian theory, the exogenous prior distributions, together with the partition of logical space with respect to which they are defined (that is, the hypothesis-space chosen), merely show where that indeterminacy is located.

3.h | Updating Rules

There is probably no other controversial topic in the Bayesian theory on which the logical perspective casts as much illumination as that of so-called *updating rules*, and in particular that of the rule of *Bayesian conditionalisation*. This rule is still widely regarded as a fundamental principle on a par with the probability axioms themselves, and a core feature of the Bayesian theory. We shall deny that it deserves this status, but at the same time show that *under suitable conditions*, which would be fulfilled in the sorts of circumstances in which updating would be typically applied, the rule is valid.

An updating rule tells 'Bayesian agents' how to adjust their belief function globally when they acquire new evidence (the reason for putting 'Bayesian agents' in quotes is because, as we have said earlier, we do not think this is a theory of agents, Bayesian or otherwise; it is a *logic*). There must, according to a widespread view, be such a rule for doing this, otherwise there could be no principled account of 'learning from experience'. The rule in this case is this. Suppose that your probability of a just before the time t, call it $t-$, at which you its learn its truth is $P_{t-}(a)$. Your new probability for a we can therefore write as $P_t(a)$, which is presumably 1. Obviously, you will have to change at least some of your other probabilities to accommodate this change consistently: for example, it might have been that $P_{t-}(b) < 1$ where b is some logical consequence of a, but consistency demands that you now change this value too to 1. This prompts two questions: (i) are there any rules which will allow one to make the necessary changes consistently, and (ii), if so, which should we adopt? To avoid subscripts we shall henceforth write P for P_{t-} and Q for P_t.

The answer to (i) is yes, and the rule almost universally recommended in the Bayesian literature is this one, called the *rule of Bayesian conditionalisation*:

$$\text{If } P(a) > 0, \text{ set } Q(\,.\,) \text{ equal to } P(\,.\mid a) \tag{3}$$

We already know (Chapter 2.**c** (**14**) and (**15**)) that Q as defined by (3) is a finitely additive probability function and is therefore a consistent way of distributing probabilities. It is also not difficult to show that the single assignment $Q(a) = 1$ can be extended consistently in ways that do not satisfy (3). So why should (3) be adopted? There are various arguments in the literature. There is, for example, the 'obvious' reason that since $P(c \mid a)$ is your probability at time t- that c is true given a, then on learning a and nothing else your updated probability of c should be a. Also, (3) has some pleasing properties: for example, that $Q(b) = 1$ if b is entailed by a is now a *consequence* (Chapter 2**c** (**13**)); also, by (**16**), Chapter 2**c**, successively conditioning on c and d is the same as successively conditioning on d and c, and so on.

But there is also a simple, and compelling, reason why (3) should not be adopted: it is *inconsistent*. Given the 'obvious' argument for (3), and the fact that it is often advertised as itself a principle of consistency, or of 'dynamic coherence', this might seem a surprising claim. Nevertheless it is easy to prove. Suppose that there is a contingent proposition b of whose truth you are P–certain; i.e. $P(b) = 1$. Suppose also that for whatever reason (the favourite example is that you are about to take a mind-altering drug) you think it distinctly P–possible that $Q(b)$ will take some value q less than 1; that is, that $P(Q(b) = q) > 0$. Let a be '$Q(b) = q$'. It follows by the probability calculus that $P(b|a) = 1$. But suppose at the appropriate time you learn a by introspection; then $Q(b) = q$. If you conditionalise on a then you

must set $Q(b) = P(b|a) = 1$.[12] Note that there is no conceptual problem, nor any implicit appeal to 'second-order probabilities' (a frequent but erroneous claim), in a statement like '$Q(b) = q$' being in the domain of an ordinary probability function: a makes a definite factual assertion, just as much as 'The next toss of this coin will land heads'. In fact, a can be written in a formally unimpeachable way as '$X_b = q$', where $X_b(s)$ is the random variable, defined on possible states of the world, whose value at s is the value of your future probability Q of b.

Nor is it an objection that the shift to '$Q(b) = q$' is not the result of rational deliberation (whatever that may mean). Indeed, such a consideration is rather obviously beside the point, which is whether conditionalisation is a *valid* rule. And it clearly is not: the information that $Q(b) = q$ cannot be conditionalised on in the circumstances described if you are consistent. Ramsey, in an often-quoted passage, pointed out that (3) might fail 'for psychological reasons' (1926, p. 180); what he did not seem to appreciate is that it might also fail for purely logico-mathematical ones.

Invalid rules breed counterexamples. What is surprising in the case of conditionalisation is that nobody seems to have realised why it, and its generalisation to Jeffrey conditionalisation in the case of an exogenous shift on any finite partition, is anything other than a completely arbitrary rule when expressed unconditionally. Why *should* I unconditionally make $Q(\ .\ )$ equal $P(\ .\ |a)$ when $P(a)$ shifts to $Q(a) = 1$? Despite the 'obvious' argument given above, there is actually nothing in the meaning of $P(\ .\ |a)$ that tells me I should: for $P(c|a)$ is my probability, *given by function* P, of c on the assumption that a is true. *Given that for reasons of consistency I now have to reject* P, to say that my probability of c should be equal to $P(c|a)$ clearly begs the question. Nor do the motley jumble of 'justifications' in the literature for the rule succeed in doing anything else. The most widely accepted of these is a so-called dynamic Dutch Book argument, due to Teller (1973), who attributes it to David Lewis. This certainly shows that you would be foolish to announce a policy in

[12] Thus this is also a counterexample to the so-called 'Reflection Principle', which says that $P(\ .\ |Q(\ .\ ) = x) = x$, $0 \leq x \leq 1$. Intuitively absurd, it has attracted a disproportionate amount of discussion.

advance of learning a for setting $Q(c)$ at a different value to $P(c \mid a)$, and then give or accept bets at both betting quotients, but the folly is your willingness to bet in that way, not the updating policy. Thus we see an interesting difference between 'dynamic' and 'synchronic' coherence: the former is entirely without probative significance, while the latter is merely a corollary of deeper principles which do not rely on considerations of financial prudence for justification.

Another alleged justification is that the updating function given by Jeffrey's rule, and hence ordinary conditionalisation which is a special case, represents the smallest departure from the initial distribution consistent with the shifted values on the partition, and hence should be the one chosen. Even if the antecedent claim is true, and that turns out entirely to depend on the appropriate choice of metric in function space, the conclusion simply begs the question. An analogy with another putative (spurious) deductive 'updating rule' is illuminating. Suppose at time t you accept the conditional $a \rightarrow b$ and at $t + 1$ you learn (and hence accept) a. The putative rule, which we might humorously call 'dynamic modus ponens', says that if by $t + 1$ you have learned nothing other than a you should accept b. Clearly the inference is not deductively valid, since (to use the terminology of Gabbay 1994) the statements have different labels. Indeed, it is easy, just as it was in the probabilistic case, to exhibit choices of a and b which actually make this spurious 'rule' yield an inconsistency. For example, a might be the negation $\sim b$ of b, so that $a \rightarrow b$ is just a convoluted way of asserting $\sim b$ (assuming that it is the material conditional being used here), and learning a then in effect means learning that $a \rightarrow b$ is false. In other words, learning a in this example undermines the previously accepted conditional.

A similar undermining happens in the probabilistic counterexample above: the initial conditional probability of 1 of b given $Q(b) = q$ is undermined by learning the latter proposition, since the conditional probability should then clearly change to q. This is intuitively obvious, but in support we can cite the fact that a bet on b conditional on a, i.e. on $Q(b) = q$, with unit stake and where the betting quotients are determined by the function Q, has the following payoff table:

b	$Q(b) = q$	
T	T	$1 - q$
F	T	$-q$
F		0

This is of course the table of a bet on b given $Q(b) = q$ with stake 1 and betting quotient q.

We now have an explanation of why conditionalisation should in suitable circumstances seem compelling. Those circumstances, just as with 'dynamic modus ponens', consist in the preservation of the relevant conditional valuations. The following probabilistic analogue of a version of modus ponens is clearly valid (to obtain the deductive rule replace $b \mid a$ by $a \rightarrow b$, let Q and P be valuations with q in $\{0,1\}$ instead of $[0,1]$):

$$\frac{Q(a) = 1 \qquad Q(b \mid a) = q}{Q(b) = q}$$

whence we obtain this conditional form of conditionalisation:

$$\frac{Q(a) = 1 \qquad Q(b \mid a) = P(b \mid a)}{Q(b) = P(b \mid a)} \tag{4}$$

But (4)'s validity is easily seen to be derived from the ordinary, 'synchronic', probability axioms. Conditionalisation is thus not a new principle to be added to the probability axioms as a 'dynamic' supplement—indeed, that way, as we saw, lies actual inconsistency—but a derived rule whose conditions of applicability are given by the standard axioms together with *assumptions* which on any given occasion may or may not be reasonable, about relevant conditional probabilities. As to a possible objection that the Bayesian methodology relies crucially on conditionalisation, the simple answer is that nothing of value can possibly be lost by recognising that the limits within which any rule is valid.

The same conditions of validity, namely the maintaining of the relevant conditional probabilities pre and post the exogenous shift, are required for Jeffrey conditionalisation, or 'probability kinematics' as Jeffrey himself calls it. When an exogenous shift

from $P(a)$ to $Q(a)$ occurs on a proposition a, Jeffrey's rule is that your new function Q should be defined by

$$Q(\,.\,) = P(\,.\mid a)Q(a) + P(\,.\mid \sim a)Q(\sim a)$$

where $Q(\sim a)$ is of course $1\text{-}Q(a)$.[13] The equation above generalises to arbitrary finite partitions. It is well known that the necessary and sufficient condition for the validity of Jeffrey's rule (its validity relative to the 'synchronic' probability axioms) is that the following equations hold: $Q(\,.\mid a) = P(\,.\mid a)$, $Q(\,.\mid \sim a) = P(\,.\mid \sim a)$. Jeffrey conditionalisation, despite alleged justificatory credentials ranging from maximum-information principles (Chapter **9.d**) to commutative diagrams,[14] is no more absolutely valid than Bayesian conditionalisation.

3.i The Cox-Good Argument

A compelling way of showing why the probability axioms can be regarded as conditions for consistent uncertain reasoning is due independently to R.T. Cox (1946, 1961) and I.J. Good (1950), though as Cox's discussion was both earlier and more systematic it is the one we shall briefly describe here. Cox was a working physicist (he was Professor of Physics at Johns Hopkins University) who sought to set out the fundamental laws of probabilistic reasoning independently of any particular scale of measurement (1961, p. 1). To this end, he proved that any function on an algebra of propositions which obeyed these laws can always be rescaled as a probability function. To be more precise, suppose that $M(a \mid b)$ is a function where a and b are members of a set S of propositions containing logical truths, falsehoods and closed under the operations &, **v**, ~, such that M takes values in an interval I of real numbers, assigns the same values to equivalent propositions, and

[13] Jeffrey's rule generalises straightforwardly to a simultaneous shift on the members of a partition (i.e. on a set of exclusive and exhaustive propositions), but not to a shift on any finite number of logically independent propositions.

[14] See, for example, Diaconis and Zabell 1982, Williams 1980, Shore and Johnson 1980, and van Fraassen 1989, pp. 331–37.

1. $M(\sim a|c) = f(M(a|c))$
2. $M(a\&b|c) = g(M(a|b\&c), M(b|c))$

for any consistent c, where f is strictly decreasing and g is increasing in both arguments and each is sufficiently smooth, i.e. satisfy certain differentiability conditions (f must have continuous second derivatives). The nub of Cox's proof is showing that the underlying logical rules governing &, **v** *and* ~ entail that f is associative in I, i.e. that $f(x,f(y,z)) = f(f(x,y),z)$, a functional equation whose general solution is of the form $Ch(f(x,y)) = h(x).h(y)$ where h is an arbitrary increasing continuous function.[15] C is the value assigned to certainty and can be set equal to unity. Thus using the rescaling function h we obtain the so-called multiplication law of probabilities, a form of axiom (4). Using the product form for f, Cox then proved that $g(x)$ must have the form $(1 - x^m)^{1/m}$, and since the product rule is also satisfied if h is replaced with h^m, we can without loss of generality take $m = 1$ in both. From this it is a short step to showing that there is an increasing (and hence order-preserving) function $h:I \rightarrow [0,1]$ such that $hM(a|c)$ is a conditional probability $P(a|c)$ on S.[16] Defining $P'(a) = P(a|t)$ where t is a logical truth, it follows that P' obeys the unconditional probability axioms together with the rule that $P'(a \ \& \ b) = P'(a|b)P'(b)$.

To sum up: Cox has shown that for any M satisfying 1.and 2. there is a probability function which orders the propositions equivalently. This result might seem too weak to be useful: there are infinitely many different algebras with the same total orderings of their elements but which admit quite different representing probability functions. But there are additional considerations which will determine the function uniquely. As de Finetti showed, if the algebra is embedded in one permitting arbitrarily many

[15] By taking logarithms the general solution could equally well be stated in the form $H(f(x,y)) = H(x) + H(y)$.

[16] An objection to Cox's argument is that the associativity of f is demonstrated only for triples (x,y,z) where $x = M(a|b\&c\&d)$, $y = M(b|c\&d)$, $z = M(c|d)$ for arbitrary a,b,c,d, which certainly do not exhaust I^3 if the domain of M is finite (Halpern 1999). A reasonable response is that Cox is considering not only actual but *possible* values of M in I for these arguments (this is implicit in the differentiability assumptions). Paris (1994, Chapter 3) gives a proof which does not assume differentiability at all.

judgments of indifference over finite sets of alternatives, that ordering will admit a unique representing finitely additive probability functions (1937, p. 101; Savage employs the same device in his 1954, pp. 38–39).

Notice how formally similar these rules are to the Boolean valuation rules for propositional logic: they tell you that the values (probability-values, truth-values) on two primitive Boolean compounds depend functionally on the values on the components. In the deductive case, since there are only two values in play, it is much easier for the rules to tell you exactly what the dependence looks like. In the probabilistic case there is a continuum of values. Nevertheless the negation case is very simple: the value on a negation is a smooth increasing function of the value on the negated proposition. The deductive rule for conjunction assumes that a Boolean valuation V, considered formally as a two-valued probability function, entails independence of the conjuncts: we have $V(A\&B) = V(A)V(B)$. In the probabilistic case we must allow for the possibility of dependence, and Cox's axiom 2. does so in the simplest possible way, by merely requiring that the joint 'probability' of A and B given consistent C depends only on the 'probability' of A given B and C and that of B given C.

It is indeed a profound result. We also have an answer to why probability functions should be chosen as the canonical representatives of belief functions satisfying 1.and 2.: because of the facts that additive functions have greater computational simplicity and that they give a common scale for objective and subjective probabilities. The proof of Cox's result is not at all a trivial matter, however, and for this reason among others we have chosen to develop the argument for the probability axioms in terms of consistent distributions of fair betting quotients: these are more familiar objects, and the proof that they generate the probability axioms is elementary by comparison. That it requires a bit more in the way of assumptions is the price paid.

3.j Exchangeability

We end this chapter by returning to the discussion which started it. We have claimed that there are two distinct interpretations of

the probability calculus, one as the formal theory of objective probability, which plays a fundamental role in statistics, and more generally in the natural, social and biological sciences, the other as a system of consistency constraints on the distribution of personal probabilities. We have said quite a lot about the latter compared with the former, because it is with epistemic probability that we principally concerned in this book. But we shall conclude by looking briefly at a very influential case made by de Finetti that a separate theory of objective probability is redundant (he also thought the idea of there being such things as objective probabilities reprehensibly metaphysical).

To discuss his arguments we first need to look at what he called *exchangeable random quantities*. Suppose that we are considering a possibility space W consisting of all denumerably infinite sequences of 0s and 1s. We can think of this as the (ideal) set of all possible outcomes in a sequence of indefinitely repeated observations, where 1 and 0 record respectively the observation of some specific characteristic. Suppose that P is a Bayesian personal probability function which assigns definite values to all propositions of the form $X_i = x_i$, to be read as 'the ith outcome is x_i', where x_i is 0 or 1. These variables are said to be *exchangeable* (with respect to P) if the P-probability of any finite conjunction $X_{i1} = x_{i1} \,\&\, \ldots \,\&\, X_{ik} = x_{ik}$ remains the same for any permutation of the x_{ij}.

De Finetti regarded the notion of an exchangeability as providing a solution to what he saw as two outstanding problems: one is the problem of induction, and the other is that of explaining why a theory of objective probability is redundant. His solution of the first is based on the second, and that stems from a celebrated theorem de Finetti proved about sequences of exchangeable variables. He showed that if a sequence like the X_i above is exchangeable then the P-probability of any sequence of n of them taking the value 1 r times and 0 the remaining n – r, in some given order, is independent of that order and equal to

$$\int z^r(1-z)^{n-r}\,dF(z) \qquad (1)$$

where the integration is from minus infinity to plus infinity, and $F(z)$ is a distribution function, uniquely determined by the

P-values on the X_i, of a random variable Z equal to the limit, where that is defined, of the random variables $Y_m = m^{-1}SX_i$.

But as we saw earlier, (1) is also the expression you get for the value of the (epistemic) probability r of the X_i taking the value 1 if you believe (i) that there is an objective, but unknown, probability p of a 1 at any point in the sequence, where F defines the epistemic probability distribution over p, over the closed unit interval [0,1], (ii) that the X_i are independent relative to this unknown probability, and (iii) you evaluate the epistemic probability of r 1s and n-r 0s to be equal to the unknown probability of the same event. The significance of de Finetti's result (1) is that we obtain formally the same explanation of the apparent convergence of the relative frequencies without needing to appeal either to the existence of an objective probability (which he repudiated for reasons based on a personal positivistic philosophy) or to the additional hypothesis h of independence with constant p. As we observed, de Finetti believed that his theorem reveals the hypothesis of independence and constant probability with respect to an objective probability to be merely redundant metaphysical baggage, the phenomena they purport to explain being merely a consequence (for consistent reasoners) of a prior judgment that certain events are exchangeable—a judgment that amounts to no more than saying that in your opinion they betray no obvious causal dependence, a very mild judgment indeed.

Or is it? We claim that it is not, and that the independence assumption apparently eschewed in (1) is nonetheless implicitly present. Consider two finite sequences of length $2n$: one, s_1, has the form

01010101010 . . . 01

while the other, s_2, is some fairly disorderly sequence of 0s and 1s, also having n 0s and n 1s, and ending with a 1. Let s_3 be a third sequence, obtained by eliminating the terminal 1 of s_2. Assume again that all the sequences, i.e. all the variables X_i, are exchangeable. By the probability calculus

$$P(s_1) = P(1 \mid 010101 \ldots 0)P(010101 \ldots 0)$$

And

$$P(s_2) = P(1 \mid s_3)P(s_3).$$

By exchangeability

$$P(010101 \ldots 1) = P(s_3)$$

And hence

$$P(1 \mid s_3) = P(1|010101 \ldots 0) \qquad (3)$$

But (3) holds for all n, however large. Were we to deny that the X_i are independent there could surely be no reason for accepting (3) for large values of n; we should on the contrary expect that the right hand side would move to the neighbourhood of 1 and the left hand side to stay in the region of 1/2 or thereabouts. We should certainly not expect equality (this example is in Good 1969, p. 21).

This is not a proof that exchangeability deductively implies independence (which is certainly not true), but it is a powerful reason supposing that exchangeability assumptions would not be made relative to repeated trials *unless* there was already a belief that the variables were independent. It is noteworthy that there is a proof, due to Spielman (1976), that if the X_i are all exchangeable according to your personal probability measure (assumed countably additive) then you must believe with probability one with respect to that same measure that they constitute a von Mises *Kollectiv*.

CHAPTER 4
Bayesian Induction: Deterministic Theories

Philosophers of science have traditionally concentrated mainly on deterministic hypotheses, leaving statisticians to discuss how statistical, or non-deterministic theories should be assessed. Accordingly, a large part of what naturally belongs to philosophy of science is normally treated as a branch of statistics, going under the heading 'statistical inference'. It is not surprising therefore, that philosophers and statisticians have developed distinct methods for their different purposes. We shall follow the tradition of dealing separately with deterministic and statistical theories. As will become apparent however, we regard this separation as artificial and shall in due course explain how Bayesian principles provide a unified scientific method.

4.a Bayesian Confirmation

Information gathered in the course of observation is often considered to have a bearing on the merits of a theory or hypothesis (we use the terms interchangeably), either by confirming or disconfirming it. Such information may derive from casual observation or, more commonly, from experiments deliberately contrived with a view to obtaining relevant evidence. The idea that observations may count as evidence either for or against a theory, or be neutral towards it, is at the heart of scientific reasoning, and the Bayesian approach must start with a suitable understanding of these concepts.

As we have described, a very natural one is at hand, for if $P(h)$ measures your belief in a hypothesis when you do not know the evidence, and $P(h \mid e)$ is the corresponding measure when you do, e strengthens your belief in h or, we may say, confirms it, just in

case the second probability exceeds the first. We refer in the usual way to $P(h)$ as 'the prior probability' of h, and to $P(h \mid e)$ as the 'posterior probability' of h relative to, or in the light of e, and we adopt the following definitions:

e **confirms or supports** h just in case $P(h \mid e) > P(h)$

e **disconfirms** h just in case $P(h \mid e) < P(h)$

e **is neutral towards** h just in case $P(h \mid e) = P(h)$.

One might reasonably take $P(h \mid e) - P(h)$ as measuring the *degree* of e's support for h, though other measures, involving for example the ratios of these terms, have also been suggested,[1] but disagreements on this score need not be settled in this book. We shall, however, say that when $P(h \mid e) > P(h \mid e') > P(h)$, the first piece of evidence confirms the hypothesis more than the second does.

According to Bayes's theorem, the posterior probability of a hypothesis depends on the three factors: $P(e \mid h)$, $P(e)$ and $P(h)$. Hence, if you know these, you can determine whether or not e confirms h, and more importantly, calculate $P(h \mid e)$. In practice, the various probabilities may be known only imprecisely, but as we shall show in due course, this does not undermine Bayes's theorem as a basis for scientific inference.

The dependence of the posterior probability on these three terms is reflected in three principal aspects of scientific inference. First, other things being equal, the more probable the evidence, relative to the hypothesis, the more that hypothesis is confirmed. At one extreme, if e refutes h, then $P(e \mid h) = 0$ and so disconfirmation is at a maximum, while the greatest confirmation is given when $P(e \mid h) = 1$, which will be met in practice when h logically implies e. Statistical hypotheses admit intermediate values for $P(e \mid h)$; as we show in later chapters, the higher the value, the greater the confirmation, other things being equal.

[1] For discussions of various other measures see, for example, Good 1950, and Jeffrey 2004, pp. 29–32.

Secondly, the power of e to confirm h depends on $P(e)$, that is, on the probability of e when h is not assumed to be true. This, of course, is not the same as the probability of e when h is assumed to be false; in fact $P(e)$ is related to the latter by the formula: $P(e) = P(e \mid h)P(h) + P(e \mid \sim h)P(\sim h)$, as we showed in Chapter 2 (Theorem 12). This inverse dependence of $P(h \mid e)$ on $P(e)$ corresponds to the familiar intuition that the more surprising the evidence, the more confirmation it provides.

Thirdly, the posterior probability of a hypothesis depends on its prior probability, a dependence that is sometimes discernible in attitudes to so-called 'ad hoc' hypotheses and in the frequently expressed preference 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.

We shall, in the course of this chapter, examine each of these facets of inductive reasoning.

4.b Checking a Consequence

A characteristic pattern of scientific inference occurs when a logical consequence of a theory is shown to be false and the theory thereby refuted. As we saw, this sort of inference, with its unimpeachable logic, impressed Popper so much that he made it the centrepiece and guiding principle of his scientific philosophy. Bayesian philosophy readily accommodates the crucial features of a theory's refutation by empirical evidence. For if a hypothesis h entails a consequence e, then, as is easily shown, provided $P(h) > 0$, $P(e \mid h) = 1$ and $P(h \mid \sim e) = 0$. Interpreted in the Bayesian fashion, this means that h is maximally disconfirmed when it is refuted. Moreover, as we should expect, once a theory has been refuted, no further evidence can ever confirm it, unless the refuting evidence be revoked. For if e' is any other observation that is logically consistent with e, and if $P(h \mid \sim e)$ is zero, then so is $P(h \mid \sim e \,\&\, e')$.

Another characteristic pattern of scientific inference occurs when a logical consequence of a theory is shown to be true and the theory then regarded as confirmed. Bayes's theorem shows

why and under what circumstances a theory is confirmed by its consequences. First, it follows from the theorem that a theory is always confirmed by a logical consequence, provided neither the evidence nor the theory takes either of the extreme probability values. For if h entails e, $P(e \mid h) = 1$, so that $P(h \mid e) = \frac{P(h)}{P(e)}$. Hence, provided $0 < P(e) < 1$ and $P(h) > 0$, $P(h \mid e) > P(h)$, which means that e confirms h.

Secondly, the probability axioms tell us, correctly, that succeeding confirmations by logical consequences eventually diminish in force (Jeffreys 1961, pp. 43–44). For let $e_1, e_2, \ldots, e_n$ be a succession of logical consequences of h, then

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

As we showed earlier, if h entails all the e_i, then $P(h \mid e_1 \,\&\, \ldots \,\& e_n) \geq P(h \mid e_1 \,\&\, \ldots \,\& e_{n-1})$. It follows from the Bolzano-Weierstrass theorem that the non-decreasing sequence of posterior probabilities has a limit. Clearly, the limits, as n tends to infinity, of the two posterior probabilities in this equation are the same, *viz.* $\lim P(h \mid e_1 \,\&\, \ldots \,\& e_n) = \lim P(h \mid e_1 \,\&\, \ldots \,\& e_{n-1})$. Hence, provided that $P(h) > 0$, $P(e_n \mid e_1 \,\&\, \ldots \,\& e_{n-1})$ must tend to 1. This explains why it is not sensible to test a hypothesis indefinitely. The result does not however tell us the precise point beyond which further predictions of the hypothesis are sufficiently probable not to be worth examining, for that would require a knowledge of individuals' belief structures which logic does not supply.

A third salient feature of confirmation by a theory's consequences is that in many instances, specific categories of those consequences each have their own, limited capacity to confirm. This is an aspect of the familiar phenomenon that however often a particular experiment is repeated, its results can confirm a general theory only to a limited extent; and when an experiment's capacity to generate significant confirming evidence for the theory has been exhausted through repetition, further support is

sought from other experiments, whose outcomes are predicted by other parts of the theory.[2]

This phenomenon has a Bayesian explanation (Urbach 1981). Consider a general hypothesis h and let h_r be a substantial restriction of that hypothesis. A substantial restriction of Newton's theory might, for example, express the idea that freely falling bodies near the Earth's surface descend with constant acceleration, or that the period and length of a pendulum are related by the familiar formula. Since h entails h_r, $P(h) \leq P(h_r)$, as we showed in Chapter 2, and if h_r is much less speculative than its progenitor, it will often be much more probable.

Now consider a series of predictions that are implied by h, and which also follow from h_r. If the predictions are verified, they may confirm both theories, whose posterior probabilities are 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 yields

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

Since the maximum value of the last probability in this equation is 1, it follows that however many predictions of h_r have been verified, the posterior probability of the main theory, h, can never rise above $\frac{P(h)}{P(h_r)}$. Therefore, the prior probability of h_r determines a limit to how far evidence entailed by it can confirm h. And this explains the phenomenon under consideration, for the predictions verified by means of an experiment (that is, a procedure designed

[2] This is related to the phenomenon that the more varied a body of evidence, the greater its inductive force, which we discuss in section 4.g below.

to a specified pattern) do normally follow from and confirm a much-restricted version of the predicting theory.

The arguments and explanations in this section rely on the possibility that evidence already accumulated from an experiment can increase the probability that further performances of the experiment will produce similar results. Such a possibility was denied by Popper and by his supporters, on the grounds that the probabilities involved are not objective. How then do they explain the fact, familiar to every scientist, that repeating some experiment indefinitely (or usually, more than a very few times) is pointless? Musgrave (1975) attempted an explanation. He argued that after a certain (unspecified) number of repetitions, the scientist should form a generalization to the effect that the experiment will always yield a result that is similar, in certain respects to those results already obtained, and that this generalization should then be entered into 'background knowledge'. Relative to the newly augmented background knowledge, the experiment is certain to produce the same result when it is next performed as it did before. Musgrave then appealed to the putative principle, which we discuss in the next section, that evidence confirms a hypothesis in proportion to the difference between its probability relative to the hypothesis plus background knowledge and its probability relative to background knowledge alone, that is, to $P(e \mid h \,\&\, b) - P(e \mid b)$, and inferred that even if the experiment did produce the expected result when next conducted, the hypothesis would receive no new confirmation.

A number of decisive objections can be raised against this account. First, as we show in the next section, although it forms part of the Bayesian account and seems to be a feature of science that confirmation depends in its degree upon the probability of the evidence, that principle has no basis in Popperian methodology. Popper simply invoked it ad hoc. Secondly, Musgrave's suggestion takes no account of the fact that particular experimental results may be generalized in infinitely many ways. This is a substantial objection since different generalizations give rise to different implications about future experimental outcomes. So Musgrave's explanation calls for a rule that would guide the scientist to a particular and appropriate generalization; but we cannot see how appropriateness could be

defined or such a rule possibly justified within the limitations of Popperian philosophy. Finally, the decision to designate the generalization background *knowledge*, with the effect that has on our evaluation of other theories and on our future conduct regarding for example, whether or not to repeat certain experiments, is comprehensible only if we have invested some confidence in the generalization. But then this Popperian account tacitly invokes the same kind of inductive notion as it was designed to avoid. The fact is that the phenomena concerning the confirming power of experiments and their repetitions are essentially inductive and are beyond the reach of anti-inductivist methodologies such as Popper's.

4.c | The Probability of the Evidence

In the Bayesian account, confirmation occurs when the posterior probability of a hypothesis exceeds its prior probability, and the greater the difference, the greater the confirmation. Now Bayes's theorem may be expressed in the following ways:

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

We see that the evidential force of e is entirely expressed by the ratio $\dfrac{P(e \mid \sim h)}{P(e \mid h)}$, known as the *Bayes factor*. The smaller this factor, that is to say, the more probable the evidence if the hypothesis is true than if it is false, the greater is the confirmation. In the deterministic case, where h entails e, so that $P(e \mid h) = 1$, confirmation depends inversely on $P(e)$ or $P(e \mid \sim h)$; this fact is 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 some time and you do, your faith in his powers of precognition would not be much enhanced: you would probably continue to regard his predictions as simply guesswork. But if the prediction also gave you

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 nicely with an incident in Shakespeare's *Macbeth*. The three witches, using their special brand of divination, tell Macbeth that he will soon become both Thane of Cawdor and King of Scotland. Macbeth finds these two predictions incredible:

> 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 shortly after making this declaration, he learns that the Thane of Cawdor prospered no longer, was in fact condemned to death, and that he, Macbeth, had succeeded to the title, whereupon, his attitude to the witches' powers of foresight alters entirely, and he comes to believe their other predictions.

Charles Babbage (1827), the celebrated polymath and 'father of computing', examined numerous logarithmic tables published over two centuries in various parts of the world, with a view to determining whether they derived from a common source or had been worked out independently. He found the same six errors in all but two and drew the "irresistible" conclusion that the tables containing those errors had been copied from a single original. As Jevons (1874, pp. 278–79) pointed out, the force of this conclusion springs from the fact that if the tables originated from the same source, then it is practically certain that an error in one will be reproduced in the others, but if they did not, the probability of errors being duplicated is minuscule. Such reasoning is so compelling that compilers of mathematical tables regularly protect their copyrights by purposely incorporating some minor errors "as a trap for would-be plagiarists" (L.J. Comrie)[3]; and cartographers do the same.

The inverse relationship between the probability of evidence and its confirming power is a simple and direct consequence of

[3] This is quoted in Bowden 1953, p. 4.

Bayesian theory. On the other hand, methodologies that eschew probabilistic evaluations of hypotheses, in the interests of objectivity, seem constitutionally unable to account for the phenomenon. Popper (1959a, appendix *ix) recognized the need to provide such an account and rose to the challenge. First, he conceded that, in regard to confirmation, the significant quantities are $P(e \mid h)$ and $P(e)$; he then measured the amount of confirmation or "corroboration" which e confers on h by the difference between those quantities. But Popper never said explicitly what he meant by the probability of evidence. He could not allow it a subjective connotation without compromising the intended objectivist quality of his methodology, yet he never worked out what objective significance the term could have. His writings suggest he had in mind some purely logical notion of probability, but neither he nor anyone else has managed to give an adequate account of logical probability. Secondly, Popper never satisfactorily justified his claim that hypotheses benefit in any epistemic sense from improbable evidence; indeed, the idea has been closely examined by philosophers and is generally regarded as indefensible within the Popperian scheme. (See Chapter 1, above, and, for example, Howson 2000, and 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 shown in Chapter 9, where some other criticisms of the Bayesian approach are rebutted.

4.d The Ravens Paradox

The Bayesian position that confirmation is a matter of degree, determined by Bayes's theorem, scotches a famous puzzle, first posed by Hempel (1945), known as the *Paradox of Confirmation* or sometimes as the *Ravens Paradox*. It was called a paradox because its premises seemed extremely plausible, despite their supposedly counter-intuitive consequences, and the reference to ravens stems from the paradigm hypothesis, 'All ravens are black', that is frequently used to present the problem. The alleged

difficulty arises from the following assumptions about confirmation. (RB will signify the proposition that a certain object is black and a raven, and $\overline{R}\overline{B}$ that it is neither black nor a raven.)

1. Hypotheses of the form 'All Rs are B' are confirmed by evidence of something that is both R and B. (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.*)

Now, by the Nicod Condition, 'All non-Bs are non-Rs' is confirmed by $\overline{R}\overline{B}$; and by the Equivalence Condition, so is 'All Rs are B', since the two generalizations are logically equivalent. Many philosophers regard this consequence as blatantly false, since it says that you can confirm the hypothesis that all ravens are black by observing a non-black non-raven, say, a white lie or a red herring. This seems to suggest that you could investigate that and other similar generalizations *just as well* by examining objects on your desk as by studying ravens on the wing. But that would be a non sequitur. For the fact that RB and $\overline{R}\overline{B}$ both confirm a hypothesis does not mean that they do so with equal force. And once it is recognized that confirmation is a matter of degree, the conclusion ceases to be counter-intuitive, because it is compatible with $\overline{R}\overline{B}$ confirming 'All Rs are B', but to a negligible degree. This simple point constitutes the Bayesian solution to the problem.

But a Bayesian analysis can take the matter further, first of all, by demonstrating that in the case of the paradigm hypothesis, data of the form $\overline{R}\overline{B}$ do in fact confirm to a negligible degree; secondly, by showing that Nicod's condition is not valid as a universal principle of confirmation. Consider the first point. The impact of the two data on h, 'All ravens are black', is given as follows:

$$\frac{P(h \mid RB)}{P(h)} = \frac{P(RB \mid h)}{P(RB)} \;\&\; \frac{P(h \mid \overline{R}\overline{B})}{P(h)} = \frac{P(\overline{R}\overline{B} \mid h)}{P(\overline{R}\overline{B})}.$$

These expressions can be simplified. First, $P(RB \mid h) = P(B \mid h \;\&\; R)P(R \mid h) = P(R \mid h) = P(R)$. We arrived at the last equality by assuming that whether some arbitrary object is a raven is independent of the truth of h, which seems plausible to us, at

any rate as a close approximation, though Horwich (1982, p. 59) thinks it lacks plausibility.[4] By parallel reasoning, $P(\overline{R}\overline{B} \mid h) = P(\overline{B} \mid h) = P(\overline{B})$. Also, $P(RB) = P(B \mid R)P(R)$, and $P(B \mid R) = \Sigma P(B \mid R \;\&\; \theta)P(\theta \mid R) = \Sigma P(B \mid R \;\&\; \theta)P(\theta)$, where θ represents possible values of the proportion of ravens that are black (h says that $\theta = 1$), and assuming independence between θ and R. Finally, $P(B \mid R \;\&\; \theta) = \theta$, for if the proportion of ravens in the universe that are black is θ, the probability of a randomly selected raven being black is also θ.

Combining all these considerations with Bayes's theorem yields:

$$\frac{P(h \mid RB)}{P(h)} = \frac{1}{\Sigma\theta P(\theta)} \;\&\; \frac{P(h \mid \overline{R}\overline{B})}{P(h)} = \frac{1}{P(\overline{R} \mid \overline{B})}.$$

According to the first of these equations, the ratio of the posterior to the prior probabilities of h is inversely proportional to $\Sigma\theta P(\theta)$. This means, for example, that if it were initially very probable that all or virtually all ravens are black, then $\Sigma\theta P(\theta)$ would be large and RB would confirm h rather little. While if it were initially relatively probable that most ravens are not black, the confirmation could be substantial. Intermediate degrees of uncertainty regarding θ would bring their own levels of confirmation to h.

The second equation refers to the confirmation to be derived from the observation of a non-black non-raven, and here the crucial probability term is $P(\overline{R} \mid \overline{B})$. Now presumably there are vastly more non-black things in the universe than ravens. So even if we felt certain that no ravens are black, the probability of some object about which we know nothing, except that it is not black, being a non-raven must be very high, practically 1. Hence, $P(h \mid \overline{R}\overline{B}) = (1 - \varepsilon)P(h)$, where ε is a very small positive number;

[4] Vranas (2004) interprets the assumption as asserting that whether some arbitrary object is a raven "should" be independent of h, and he criticizes this and other Bayesian accounts for depending upon a claim for which, he says, there can be no reasoned defence. But our argument does not need such a strong assumption. Our position is merely that in this particular case, our and, we suspect, most other people's personal probabilities are such that independence obtains.

therefore, observing that some object is neither a raven nor black provides correspondingly little confirmation for h.[5]

A Bayesian analysis necessarily retains the Equivalence Condition but gives only qualified backing to the Nicod Condition, for it anticipates circumstances in which the condition fails. For instance, suppose the hypothesis under examination is 'All grasshoppers are located outside the County of Yorkshire'. One of these creatures appearing just beyond the county border is an instance of the generalization and, according to Nicod, confirms it. But it might be more reasonably argued that since there are no border controls or other obstacles to restrict the movement of grasshoppers in that area, the observation of one on the edge of the county but outside it increases the probability that some others have actually crossed over, and hence, contrary to Nicod, it undermines the hypothesis. In Bayesian terms, this is a case where, relative to background information, the probability of some datum is reduced by a hypothesis—that is, $P(e \mid h) < P(e)$—which is thereby disconfirmed—that is, $P(h \mid e) < P(h)$.[6] This example is adapted from Swinburne 1971, though the idea seems to originate with Good 1961.

Another, more striking case where Nicod's Condition breaks down was invented by Rosenkrantz (1977, p. 35). Three people leave a party, each with a hat. The hypothesis that none of the three has his own hat is confirmed, according to Nicod, by the observation that person 1 has person 2's hat and by the observation that person 2 has person 1's hat. But since the hypothesis concerns only three, particular people, the second observation must *refute* the hypothesis, not confirm it.

Our grasshopper example may also be used to show that instances of the type $\bar{R}B$ can sometimes confirm 'All Rs are B'. Imagine that an object that looks for all the world like a grasshopper had been 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 after all would be relatively unlikely unless the grasshopper hypothesis were true (hence, $P(e) < P(e \mid h)$);

[5] The account given here is substantially similar to Mackie's, 1963.

[6] This example is adapted from Swinburne 1971, though the idea seems to originate with Good 1961.

so it would confirm that hypothesis. If the deceptively grasshopper-like object were discovered within the county, the same conclusion would follow for this $\bar{R}\bar{B}$ instance.

Horwich (1982, p. 58) has argued that the ravens hypothesis could be differently confirmed depending on how the black raven was chosen, either by randomly selecting an object from the population of ravens, or by restricting the selection to the population of black objects. Korb (1994) provides a convincing demonstration of this, which we discuss in a related context in Chapter 8.

We do not accept Horwich's argument for his conclusion. Denoting a black raven either R^*B or RB^*, depending on whether it was discovered by the first selection process or the second, he claims that evidence of the former kind always confirms more, because only it subjects the raven hypothesis to the risk of falsification. But this conflates the process of collecting evidence, which may indeed expose the hypothesis to different risks of refutation, with the evidence itself, which either does or does not refute the hypothesis, and in the present case it does not.

Our conclusions are, first, that the so-called paradox of the ravens is not in fact problematic; secondly, that of the two conditions of confirmation that generated it only the Equivalence Condition is acceptable; and thirdly, that Bayesian theory explains why.

4.e The Duhem Problem

The Duhem (sometimes called the Duhem-Quine) problem arises with philosophies of science of the type associated with Popper, which emphasize the power of certain evidence to refute a theory. According to Popper, falsifiability is the feature of a theory which makes it scientific. "Statements or systems of statements," he said, "in order to be ranked as scientific, must be capable of conflicting with possible, or conceivable, observations" (1963, p. 39). And claiming to apply this criterion, he judged Einstein's gravitational theory scientific and Freud's psychology not. The term 'scientific' carries a strong flavour of commendation, which is, however, misleading in this context. For Popper could never demonstrate a link between his concept

of scientificness and epistemic or inductive merit: a theory that is scientific in Popper's sense is not necessarily true, or probably true, nor can it be said either definitely or even probably to lead to the truth. There is little alternative then, in our judgment, to regarding Popper's demarcation between scientific and unscientific statements as without normative significance, but as a claim about the content and character of what is ordinarily termed science.

Yet as an attempt to understand the practice of science, Popper's ideas bear little fruit. First of all, the claim that scientific theories are falsifiable by "possible, or conceivable, observations" raises a difficulty, because an observation can only falsify a theory (in other words conclusively demonstrate its falsity) if it is itself conclusively certain. Yet as Popper himself appreciated, no observations fall into this category; they are all fallible. But unwilling to concede degrees of fallibility or anything of the kind, Popper took the view that observation reports that are admitted as evidence "are accepted as the result of a decision or agreement; and to that extent they are *conventions*" (1959a, p. 106; our italics). It is unclear to what psychological attitude such acceptance corresponds, but whatever it is, Popper's view pulls the rug from under his own philosophy, since it implies that no theory can really be falsified by evidence. Every 'falsification' is merely a convention or decision: "From a logical point of view, the testing of a theory depends upon basic statements whose acceptance or rejection, in its turn, depends upon our *decisions*. Thus it is *decisions* which settle the fate of theories" (1959a, p. 108).

Watkins was one of those who saw that the Popperian position could not rest on this arbitrary basis, and he attempted to shore it up by arguing that some infallibly true observation statements do in fact exist. He agreed that a statement like 'the hand on this dial is pointing to the numeral 6' is fallible, since it is possible, however unlikely, that the person reporting the observation mistook the position of the hand. But he claimed that introspective perceptual reports, such as 'in my visual field there is now a silvery crescent against a dark blue background', "may rightly be regarded by their authors when they make them as infallibly true" (1984, pp. 79 and 248). But in our opinion Watkins was wrong, and the statements he regarded as infallible are open to the same sceptical

doubts as any other observational report. We can illustrate this through the above example: clearly it is possible, though admittedly not very probable, that the introspector has misremembered and mistaken the shape he usually describes as a crescent, or the sensation he usually receives on reporting blue and silvery images. These and other similar sources of error ensure that introspective reports are not exempt from the rule that non-analytic statements are fallible.

Of course, the kinds of observation statement we have mentioned, if asserted under appropriate circumstances, would never be seriously doubted, for although they could be false, they have a force and immediacy that carries conviction: in the traditional phrase, they are 'morally certain'. But if they are merely indubitable, then whether or not a theory is regarded as refuted by observational data rests ultimately on a subjective feeling of certainty, a fact that punctures the objectivist pretensions of Popperian philosophy.

A second objection to Popper's falsifiability criterion, and the one upon which we shall focus for its more general interest, is that it deems unscientific most of those theories that are usually judged science's greatest achievements. This is the chief aspect of the well-known criticisms advanced by Polanyi (1962), Kuhn (1970), and Lakatos (1970), amongst others, but based on the arguments of Duhem (1905). They pointed out that notable theories of science are typically unfalsifiable by observation statements, because they only make empirical predictions in association with certain auxiliary theories. Should any such prediction turn out to be false, logic does not compel us to regard the principal theory as untrue, since the error may lie in one or more of the auxiliaries. Indeed, there are many occasions in the history of science when an important theory led to a false prediction but was not itself significantly impugned thereby. The problem that Duhem posed was this: *when several distinct theories are involved in deriving a false prediction, which of them should be regarded as false?*

Lakatos and Kuhn on the Duhem Problem

Lakatos and Kuhn both investigated scientific responses to anomalies and were impressed by the tendency they observed for

the benefit of the doubt persistently to be given to particular, especially fundamental theories, and for one or more of the auxiliary theories regularly to be blamed for any false prediction. Lakatos drew from this observation the lesson that science of the most significant kind usually proceeds in what he called scientific research programmes, each comprising a central, or 'hard core', theory, and a so-called 'protective belt' of auxiliary theories. During the lifetime of a research programme, these elements are combined to yield empirical predictions, which are then experimentally checked; and if they turn out to be false, the auxiliary hypotheses act as a protective shield, as it were, for the hard core, and take the brunt of the refutation. A research programme is also characterised by a set of heuristic rules by which it develops new auxiliary hypotheses and extends into new areas. Lakatos regarded Newtonian physics as an example of a research programme, the three laws of mechanics and the law of gravitation comprising the hard core, and various optical theories, propositions about the natures and dispositions of the planets, and so forth, being the protective belt.

Kuhn's theory is similar to the methodology we have just outlined and probably inspired it in part. Broadly speaking, Kuhn's 'paradigm' is the equivalent of a scientific research programme, though his idea is developed in less detail.

Lakatos, following Popper, also added a normative element, something that Kuhn deliberately avoided. He held that it was perfectly all right to treat the hard core systematically as the innocent party in a refutation, provided the research programme occasionally leads to successful "novel" predictions or to successful, "non-ad hoc" explanations of existing data. Lakatos called such programmes "progressive."

> The sophisticated falsificationist [which Lakatos counted himself] . . . sees nothing wrong with a group of brilliant scientists conspiring to pack everything they can into their favourite research programme . . . with a sacred hard core. As long as their genius—and luck—enables 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 research programme persistently produces false predictions, or if its explanations are habitually ad hoc, Lakatos called it "degenerating." The notion of an ad hoc explanation—briefly, one that does not produce new and verified predictions—is central to attempts by the Popperian school to deal with the Duhem problem and we discuss it in greater detail below, in section **g**. In appraising research programmes, Lakatos employed the tendentious terms 'progressive' and 'degenerating', but he never succeeded in substantiating their normative intimations, 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 were well regarded by scientists, while degenerating ones were distrusted and eventually dropped.

This last claim, it seems to us, contains a measure of truth, as evidenced by case studies in the history of science, such as those in Howson 1976. But although Lakatos and Kuhn identified and described an important aspect of scientific work, they could not explain it or rationalize it. So, for example, Lakatos did not say why a research programme's occasional predictive success could compensate for numerous failures, nor did he specify how many such successes are needed to convert a degenerating programme into a progressive one, beyond remarking that they should occur "now and then".

Lakatos was also unable to explain why certain theories are raised to the privileged status of hard core in a research programme while others are left to their own devices. His writings give the impression that the scientist is free to decide the question at will, by "methodological fiat", as he says. Which suggests that it is 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 attribute all empirical difficulties to auxiliary hypotheses. This is far from being the case. For these reasons and also because of difficulties with the notion of an ad hoc hypothesis, to be discussed below, neither Kuhn's theory of paradigms nor Lakatos's so-called 'sophisticated falsificationism' are in any position to solve the Duhem problem.

The Bayesian Resolution

The questions left unanswered in the Kuhn and Lakatos methodologies are addressed and resolved, as Dorling (1979) brilliantly

showed, by referring to Bayes's theorem and considering how the individual probabilities of theories are severally altered when, as a group, they have been falsified.

We shall illustrate the argument through a historical example that Lakatos (1970, pp. 138–140; 1968, pp. 174–75) drew heavily upon. In the early nineteenth century, William Prout (1815, 1816), a medical practitioner and chemist, advanced the idea that the atomic weight of every element is a whole-number multiple of the atomic weight of hydrogen, the underlying assumption being that all matter is built up from different combinations of some basic element. Prout believed hydrogen to be that fundamental building block. Now many of the atomic weights recorded at the time were in fact more or less integral multiples of the atomic weight of hydrogen, but some deviated markedly from Prout's expectations. Yet this did not shake the strong belief he had in his hypothesis, for in such cases he blamed the methods that had been used to measure those atomic weights. Indeed, he went so far as to adjust the atomic weight of the element chlorine, relative to that of hydrogen, from the value 35.83, obtained by experiment, to 36, the nearest whole number. Thomas Thomson (1818, p. 340) responded in a similar manner when confronted with 0.829 as the measured atomic weight (relative to the atomic weight of oxygen) of the element boron, changing it to 0.875, "because it is a multiple of 0.125, which all the atoms seem to be". (Thomson erroneously took the relative atomic weights of hydrogen and oxygen as 0.125.)

Prout's reasoning relative to chlorine and Thomson's, relative to boron, can be understood in Bayesian terms as follows: Prout's hypothesis t, together with an appropriate assumption a, asserting the accuracy (within specified limits) of the measuring techniques, the purity of the chemicals employed, and so forth, implies that the ratio of the measured atomic weights of chlorine and hydrogen will approximate (to a specified degree) a whole number. In 1815 that ratio was reported as 35.83—call this the evidence e—a value judged to be incompatible with the conjunction of t and a.

The posterior and prior probabilities of t and of a are related by Bayes's theorem, as follows:

$$P(t \mid e) = \frac{P(e \mid t)P(t)}{P(e)} \quad \text{and} \quad P(a \mid e) = \frac{P(e \mid a)P(a)}{P(e)} \ .$$

To evaluate the two posterior probabilities, it is necessary to quantify the various terms on the right-hand sides of these equations.

Consider first the prior probabilities of t and of a. J.S. Stas, a distinguished Belgian chemist whose careful atomic weight measurements were highly influential, gives us reason to think that chemists of the period were firmly disposed to believe in t, recalling that "In England the hypothesis of Dr Prout was almost universally accepted as absolute truth" and that when he started investigating the subject, he himself had "had an almost absolute confidence in the exactness of Prout's principle" (1860, pp. 42 and 44).

It is less easy to ascertain how confident Prout and his contemporaries were in the methods used to measure atomic weights, but their confidence was probably not great, in view of the many clear sources of error. For instance, errors were recognised to be inherent in the careful weighings and manipulations that were required; the particular chemicals involved in the experiments to measure the atomic weights were of questionable purity; and, in those pioneer days, the structures of chemicals were rarely known with certainty.[7] These various uncertainties were reinforced by the fact that independent measurements of atomic weights, based on the transformations of different chemicals, rarely delivered identical results.[8] On the other hand, the chemists of the time must have felt that that their atomic weight measurements were more likely to be accurate than not, otherwise they would hardly have reported them.[9]

[7] The several sources of error were rehearsed by Mallet (1893).

[8] For example, Thomson (1818, p. 340) reported two independent measurements—2.998 and 2.66—for the weight, relative to the atomic weight of oxygen, of a molecule of boracic (boric) acid. He required this value in order to calculate the atomic weight of boron from the weight of the boric acid produced after the element was combusted.

[9] "I am far from flattering myself that the numbers which I shall give are all accurate; on the contrary, I have not the least doubt that many of them are still erroneous. But they constitute at least a nearer approximation to the truth than the numbers contained in the first table [which Thomson had published some years before]" (Thomson 1818, p. 339).

For these reasons, we conjecture that $P(a)$ was in the neighbourhood of 0.6 and that $P(t)$ was around 0.9, and these are the figures we shall work with. We stress that these figures and those we shall assign to other probabilities are intended chiefly to show that hypotheses that are jointly refuted by an observation, may sometimes be disconfirmed to very different degrees, so illustrating the Bayesian resolution of Duhem's problem. Nevertheless, we believe that the figures we have suggested are reasonably accurate and sufficiently so to throw light on the historical progress of Prout's hypothesis. As will become apparent, the results we obtain are not very sensitive to variations in the assumed prior probabilities.

The posterior probabilities of t and of a depend also on $P(e)$, $P(e \mid t)$, and $P(e \mid a)$. Using the theorem of total probability, the first two of these terms can be expressed as follows:

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

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

We will follow Dorling in taking t and a to be independent, viz, $P(a \mid t) = P(a)$ and hence, $P(\sim a \mid t) = P(\sim a)$. As Dorling points out (1996), this independence assumption makes the calculations simpler but is not crucial to the argument. Nevertheless, that assumption accords with many historical cases and seems clearly right here. For we put ourselves in the place of chemists of Prout's day and consider how our confidence in his hypothesis would have been affected by a knowledge that particular chemical samples were pure, that particular substances had particular molecular structures, specific gravities, and so on. It seems to us that it would not be affected at all. Bovens and Hartmann (2003, p. 111) take a different view and have objected to the assumption of independence in this context. Speaking in general terms, they allege that "experimental results are determined by a hypothesis and auxiliary theories that are often hopelessly interconnected with each other."

> And these interconnections raise havoc in assessing the value of experimental results in testing hypotheses. There is always the fear that the hypothesis and the auxiliary theory really come out of the same deceitful family and that the lies of one reinforce the lies of the other.

We do not assert that theories are never entangled in the way that Bovens and Hartmann describe, but for the reasons we have just cited, it seems to us that the present situation is very far from being a case in point.

Returning to the last equation, if we incorporate the independence assumption and take account of the fact that since the conjunction $t \,\&\, a$ is refuted by e, $P(e \mid t \,\&\, a)$ must be zero, we obtain:

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

By parallel reasoning, we may derive the results:

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

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

So, 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 calculated:

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

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

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

The first of these gives the probability of the evidence if Prout's hypothesis is not true, but if the assumptions made in calculating the atomic weight of chlorine are accurate. Certain nineteenth-century chemists thought carefully about such probabilities, and typically took a theory of random distribution of atomic weights as the alternative to Prout's hypothesis (for instance, Mallet 1880); we shall follow this. Suppose it had been established for certain that the atomic weight of chlorine lies between 35 and 36. (The final results we obtain respecting the posterior probabilities of t and of a are, incidentally, unaffected by the width of this interval.) The random-distribution theory assigns equal probabilities to the atomic weight of an element lying in any 0.01-wide band. 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 attributed 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 had been inaccurately weighed, 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 measurements were moderately likely at the time, chemists would not have considered their techniques entirely haphazard. Thus if Prout's hypothesis were true and the measurement technique imperfect, the measured atomic weights would be likely to deviate somewhat from integral values; but the greater the deviation, the less the likelihood, so the probability distribution of atomic weight measurements falling within the 35–36 interval would not be uniform, but would be more concentrated around the whole numbers.

Let us proceed with the figures we have proposed for the crucial probabilities. We note however that the absolute values of the probabilities are unimportant, for, in fact, only their relative values count in the calculation. Thus we would arrive at the same results with the weaker assumptions that $P(e \mid \sim t \ \& \ a) = P(e \mid \sim t \ \& \ \sim a) = \frac{1}{2}P(e \mid t \ \& \ \sim a)$. We now obtain:

$$P(e \mid \sim t) = 0.01 \times 0.6 + 0.01 \times 0.4 = 0.01$$

$$P(e \mid t) = 0.02 \times 0.4 = 0.008$$

$$P(e \mid a) = 0.01 \times 0.1 = 0.001$$

$$P(e) = 0.008 \times 0.9 + 0.01 \times 0.1 = 0.0082.$$

Finally, Bayes's theorem allows us to derive the posterior probabilities in which we are interested:

$P(t \mid e) = 0.878$ (Recall that $P(t) = 0.9$)

$P(a \mid e) = 0.073$ (Recall that $P(a) = 0.6$).

We see then that the evidence provided by the measured atomic weight of chlorine affects Prout's hypothesis and the set of auxiliary hypotheses very differently; for while the probability of the first is scarcely changed, that of the second is reduced to a point where it has lost all credibility.

It is true that these results depend upon certain—we have argued plausible—premises concerning initial probabilities, but this does not seriously limit their general significance, because quite substantial variations in the assumed probabilities lead to quite similar conclusions, as the reader can verify. So for example, if the prior probability of Prout's hypothesis were 0.7 rather than 0.9, the other assignments remaining unchanged, $P(t \mid e)$ would equal 0.65, and $P(a \mid e)$ would be 0.21. Thus, as before, Prout's hypothesis is still more likely to be true than false in the light of the adverse evidence, and the auxiliary assumptions are still much more likely to be false than true.

Successive pieces of adverse evidence may, however, erode the probability of a hypothesis so that eventually it becomes more likely to be false than true and loses its high scientific status. Such a process would correspond to a Lakatosian degenerating research programme or be the prelude to a Kuhnian paradigm shift. In the present case, the atomic weight of chlorine having been repeated in various, improved ways by Stas, whose laboratory skill was universally recognized, Mallet (1893, p. 45) concluded that "It may be reasonably said that probability is against the idea of any future discovery . . . ever making the value of this element agree with an integer multiple of the atomic weight of hydrogen". And in the light of this and other atomic weight measurements he regarded Prout's original idea as having been "shown by the calculus of probability to be a very improbable one". And Stas himself, who started out so very sure of its truth, reported in 1860 that he had now "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).

We conclude that Bayes's theorem provides a framework that resolves the Duhem problem, unlike the various non-probabilistic methodologies which philosophers have sought to apply to it. And the example of Prout's hypothesis, as well as others that Dorling

(1979 and 1996) has analysed, show in our view, that the Bayesian model is essentially correct.

4.f Good Data, Bad Data, and Data Too Good to Be True

Good Data

The marginal influence that an anomalous observation may exert on a theory's probability contrasts with the dramatic effect of some confirmations. For instance, if the measured atomic weight of chlorine had been a whole number, in line with Prout's hypothesis, so that $P(e \mid t \,\&\, a) = 1$ instead of 0, and if the other probability assignments remained the same, the probability of the hypothesis would shoot up from a prior of 0.9 to a posterior of 0.998. And even more striking: had the prior probability of t been 0.7, its posterior probability would have risen to 0.99.

This asymmetry between the effects of anomalous and confirming instances was emphasized by Lakatos, who regarded it as highly significant in science, and as a characteristic feature of a research programme. He maintained that a scientist involved in such a programme typically "forges ahead with almost complete disregard of 'refutations'," provided there are occasional predictive successes (1970, p. 137): the scientist is "encouraged by nature's YES, but not discouraged by its NO" (p. 135). As we have indicated, we believe there to be much truth in Lakatos's observations; the trouble, however, is that these observations are merely absorbed, without justification, into his methodology; the Bayesian methodology, on the other hand, explains why and under what circumstances the asymmetry effect is present.

Bad Data

An interesting fact that emerges from the Bayesian analysis is that a successful prediction derived from a combination of two theories does not necessarily redound to the credit of both of them,

indeed one may even be discredited. Consider Prout's hypothesis again, and suppose the atomic weight of chlorine had been determined, not in the established way, but by concentrating hard on the element while selecting a number blindly from a given range of numbers. And let us suppose that the atomic weight of chlorine is reported by this method to be a whole number. This is just what one would predict on the basis of Prout's hypothesis, if the outlandish measuring technique were accurate. But accuracy is obviously most unlikely, and it is equally obvious that the results of the technique could add little or nothing to the credibility of Prout's hypothesis. This intuition is upheld by Bayes's theorem: 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 above. And, because, as we said, a is extremely implausible, we will set $P(a)$ at, say 0.0001. It then follows that t is not significantly confirmed by e, for $P(t)$ and $P(t \mid e)$ are virtually identical.

This example shows that Leibniz was wrong to declare as a maxim that "It is the greatest commendation of a hypothesis (next to truth) if by its help predictions can be made even about phenomena or experiments not [yet] 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 may accrue very little credit when the prediction is verified. This explains why the various sensational predictions that Velikovsky drew from his theory failed to impress most serious astronomers, even when some of those predictions were to their amazement fulfilled. For instance, Velikovsky's prediction (1950, p. 351) 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 string of unsupported and implausible assumptions, for instance, that the comet in question carried hydrogen and carbon; that these had been converted to petroleum by electrical discharges supposedly generated in the violent impact with the Earth; that the comet had later evolved into the planet Venus; and some others. (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 fit a favoured hypothesis more perfectly than seems reasonable. Imagine, for instance, that Prout had advanced his hypothesis and then proceeded to report numerous atomic weights that he had himself measured, each an exact whole number. Such a result looks almost as if it was designed to impress, and just for this reason it fails to.

We may analyse this response as follows: chemists in the early nineteenth century recognized that the measuring techniques available to them were not absolutely precise in their accuracy but were subject to experimental error, and so liable to produce a certain spread of results about the true value. On this assumption, which we label a', it is extremely unlikely that numerous independent atomic weight measurements would all produce exactly whole numbers, even if Prout's hypothesis were true. So $P(e \mid t \,\&\, a')$ is extremely small, and clearly $P(e \mid \sim t \,\&\, a')$ could be no larger. Now there are many possible explanations of e, apart from those involving a', one being that the experiments were consciously or unconsciously rigged so as to appear favourable to Prout's hypothesis. If this were the only plausible alternative (and so, in effect, equivalent to $\sim a$), $P(e \mid t \,\&\, \sim a')$ would be very high, as too $P(e \mid \sim t \,\&\, \sim a)$. It follows from the equations in section **e,** 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 exactly whole numbers would be equally effective whether t was true or not; in other words,

$$P(e \mid t \,\&\, \sim a') = P(e \mid \sim t \,\&\, \sim a').$$

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 we have made, this implies that the idea that the experiments were fabricated is rendered more probable. (The above analysis is essentially due to Dorling 1996.)

A famous case of data that were alleged to be 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 of different plants producing specific kinds of offspring. For example, under certain circumstances, pea plants of a certain 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 said, "would contravene the weight of the evidence supplied in detail by . . . [Mendel's] paper as a whole". 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 that he backed up with some, rather meagre, evidence. Dobzhansky (1967, p. 1589), on the other hand, thought it "at least as plausible" that Mendel had himself discarded results that deviated much from his ideal, in the sincere belief that they were contaminated or that some other accident had befallen them. (For a comprehensive review see Edwards 1986.)

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 would explain 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 implies that if the method of eliciting atomic weights had long been established as precise and accurate, and if careful precautions had been taken against experimenter bias and deception, so that all the natural alternatives to Prout's hypothesis could be discounted, the inductive force of the data would then no longer be suspicious. Fisher, however, did not subscribe to the principle, at least, not explicitly; he believed that Mendel's results told against the genetic theory, irrespective of any alternative explanations that might be suggested. But despite this official position, Fisher did in fact, as we have just indicated, sometimes appeal to such alternatives when he formulated his argument. We refer again to Fisher's case against Mendel in the next chapter, section **b**.

4.g Ad Hoc Hypotheses

We have been discussing the circumstances in which an important scientific hypothesis, in combination with others, makes a false prediction and yet emerges with its reputation more or less intact, while one or more of the auxiliary hypotheses are largely discredited. We argued that this process necessarily calls for alternatives to the discredited hypotheses to be contemplated. Philosophers, such as Popper and Lakatos, who deny any inductive role for evidence, and who oppose, in particular, the Bayesian approach take note of the fact that scientists often do deal with particular instances of the Duhem problem by proposing alternative hypotheses; some of these philosophers have suggested certain normative rules that purport to say when such alternatives are acceptable and when they are not. Their idea is that a theory that was introduced ad hoc, that is, "for the sole purpose of saving a hypothesis seriously threatened by adverse evidence" (Hempel 1966, p. 29), is in some way inferior. The adhocness idea was largely inspired by certain types of scientific example, which appeared to endorse it, but in our view, the examples are misinterpreted and the idea badly flawed. The following are four such examples.

1 Velikovsky, in a daring book called *Worlds in Collision* that attracted a great deal of interest and controversy some years ago,

advanced the theory that the Earth has been subject at various stages in its history to cosmic disasters, through near collisions with massive comets. He claimed that one such comet passed close by our planet during the Israelites' captivity in Egypt, causing many of the remarkable events related in the Bible, such as the ten plagues and the parting of the Red Sea, before settling down as the planet Venus. Because the putative cosmic encounter rocked the entire Earth, Velikovsky expected other peoples to have recorded its consequences too, if they kept records at all. But as a matter of fact, many communities around the world failed to note anything out of the ordinary at the time, an anomaly that Velikovsky attributed to 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.

2 Dianetics is a theory that purports to analyse the causes of insanity and mental stress, which it sees as caused by the 'misfiling' of information in unsuitable locations in the brain. By re-filing these 'engrams', it claims, sanity may be restored, composure enhanced and, incidentally, the memory vastly improved. The therapy is long and expensive and few people have been through it and borne out the theory's claims. However, L. Ron Hubbard, the inventor of Dianetics, trumpeted one purported success, and exhibited this person to a large audience, saying that she had a "full and perfect recall of every moment of her life". But questions from the floor ("What did you have for breakfast on October 3rd, 1942?", "What colour is Mr Hubbard's tie?", and the like) soon demonstrated that the hapless woman had a most imperfect memory. Hubbard explained to the dwindling assembly that when she first appeared on the stage and was asked to come forward "now", the word had frozen her in "present time" and paralysed her ability to recall the past. (See Miller 1987.)

3 Investigations into the IQs of different groups of people show that the average levels of measured intelligence vary. Some

environmentalists, so-called, attribute low scores primarily to poor social and educational conditions, an explanation that ran into trouble when a large group of Inuit, leading an aimless, poor and drunken existence, were found to score very highly on IQ tests. The distinguished biologist Peter Medawar (1974), in an effort to deflect the difficulty away from the environmentalist thesis, tried to explain this unexpected 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 that was proposed in place of the refuted one seems highly unsatisfactory. It is not likely that any of them would have been advanced, save in response to particular anomalies and in order to evade the consequent difficulty, hence the label 'ad hoc'. But philosophers who attach inductive significance to adhocness recognize that the mere fact that the theory was proposed under such circumstances is not by itself grounds for condemnation. For there are examples, like the following, where a theory that was proposed for the sole purpose of dealing with an anomaly was nevertheless very successful.

4 William Herschel, in 1781, discovered the planet Uranus. Astronomers quickly sought to describe the orbit of the new planet in Newtonian terms, taking account of the perturbing influence of the other known planets, and were able to deduce predictions concerning its future positions. But discrepancies between predicted and observed positions of Uranus substantially exceeded the accepted limits of experimental error, and grew year by year. A few astronomers mooted the possibility that the fault lay with Newton's laws but the prevailing opinion was that there must be some unknown planet acting as an extra source of gravitational attraction on Uranus, which ought to be included in the Newtonian calculations. Two astronomers in particular, Adams and Le Verrier, working independently, were convinced of this and using all the known sightings of Uranus, they calculated in a mathematical *tour de force* where the hypothetical planet must be. The hypothesis was ad hoc, yet it was vindicated when careful telescopic observations as well as studies of old astronomical charts

revealed in 1846 the presence of a planet with the anticipated characteristics. The planet was later called Neptune. Newton's theory was saved, for the time being. (See Smart 1947.)

The Adhocness Criteria

Examples like the first three above have suggested to some philosophers that when a theory t, and an auxiliary hypothesis a, are jointly refuted by some evidence, e', then any replacement, of the form t & a', must not only imply e', but should also have some new, 'independent' empirical implications. And examples similar to the fourth have suggested that if the new theory satisfies this condition, then it is a particular virtue if some of the new, independent implications are verified.

These two criteria were anticipated some four hundred years ago, by the great philosopher Francis Bacon, who objected to any hypothesis that is "only fitted to and made to the measure of those particulars from which it is derived". He argued that a hypothesis should be "larger or wider" than the observations that gave rise to it and said that "we must look to see whether it confirms its largeness and wideness by indicating new particulars" (1620, I, 106). Popper (1963, p. 241) advanced the same criteria, laying it down that a "new theory should be *independently testable*. That is to say, apart from explaining all the *explicanda* which the new theory was designed to explain, it must have new and testable consequences (preferably consequences of a *new kind*)." And secondly, he said, the new theory "should pass the independent tests in question". Bacon called hypotheses that did not meet the criteria "frivolous distinctions", while Popper termed them "ad hoc".[10]

[10] The first recorded use of the term 'ad hoc' in this context in English was in 1936, in a review of a psychology book, where the reviewer criticized some explanations proffered by the book's author for certain aspects of childish behaviour:

> There s a suspicion of 'ad-hoc-ness' about the 'explanations'. 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, p. 249; our italics)

Lakatos (1970, p. 175) refined this terminology, calling a theory that failed the first requirement ad hoc_1, and one that failed the second ad hoc_2, intending these, of course, as terms of disapproval. By these criteria, the theories that Velikovsky, Medawar, and Hubbard advanced in response to anomalous data are probably ad hoc_1, for they seem to make no independent predictions, though of course a closer study of those theories might reverse that assessment. The Adams-Le Verrier hypothesis, on the other hand, is ad hoc in neither sense, because it did make new predictions, some of which were verified by telescopic sightings of Neptune. Again, philosophical and intuitive judgment coincides. Nevertheless, the adhocness criteria are unsound.

This unsoundness is evident both on apriori grounds and through counter-examples, some of which we consider now. For instance, suppose one were examining the hypothesis that a particular urn contains only white counters, and imagine an experiment in which a counter is withdrawn from the urn at random and then, after its colour has been noted, replaced; and suppose that in 10,000 repetitions of this operation 4,950, say, of the selected counters were red and the rest white. This evidence clearly refutes the initial hypothesis taken together with the various necessary auxiliary hypotheses, and it is then natural to conclude that, contrary to the original assumption, the urn contains both red and white counters in approximately equal numbers. This inference seems perfectly reasonable, and the revised hypothesis appears well justified by the evidence, *yet there is no independent evidence for it*. And if we let the urn vaporize immediately after the last counter has been inspected, no such independent evidence would be possible. So the hypothesis about the (late) urn's contents is ad $hoc_{1\,\&\,2}$; but for all that, it seems plausible and satisfactory (Howson 1984; Urbach 1991).

Speculating on the contents of an urn is but a humble form of enquiry, but there are many instances in the higher sciences which have the same import. Take the following one from the science of genetics: suppose it was initially proposed or believed that two phenotypic characteristics of a certain plant are inherited in accordance with Mendel's principles, through the agency of a pair of independently acting genes located on different chromosomes. Imagine now that plant-breeding experiments throw up a surprising number of plants carrying both phenotypes, so that the original hypothesis of independence is rejected in favour of the idea that the genes are linked on the same chromosome. Again, the revised theory would be strongly confirmed, and established as acceptable merely on the evidence that discredited its predecessor, without any further, independent evidence. (Fisher 1970, Chapter IX, presented an example of this sort.)

The history of the discovery of Neptune, which we have already discussed, illustrates the same point. Adams estimated the mass of the hypothetical planet and the elements of its orbit by the mathematical technique of least squares applied to all the positional observations available on Uranus. Adams's hypothesis fitted these observations so well that *even before Neptune had been sighted through the telescope or detected on astronomical charts*, its existence was contemplated with the greatest confidence by the leading astronomers of the day. For instance, in his retirement address as president of the British Association, Sir John Herschel, after remarking that the previous year had seen the discovery of a minor planet, went on: "It has done more. It has given us the probable prospect of the discovery of another. We see it as Columbus saw America from the shores of Spain. Its movements have been felt, trembling along the far-reaching line of our analysis, *with a certainty hardly inferior to that of ocular demonstration*". And the Astronomer Royal, Sir George Airy, who was initially inclined to believe that the problem with Uranus would be resolved by introducing a slight adjustment to the Inverse-Square law, spoke of "*the extreme probability* of now discovering a new planet in a very short time" (quoted by Smart, p. 61; our italics). Neptune was indeed discovered within a very short time.

There is a more general objection to the idea that hypotheses are unacceptable if they are ad hoc. Imagine a scientist who is interested in the conjunction of the hypotheses t & a, whose implication e can be checked in an experiment. The experiment is performed with the result e', incompatible with e, and the scientist ventures a new theory t & a', which is consistent with the observations. And suppose that either no new predictions follow or none has been confirmed, so that the new theory is ad hoc.

Imagine that another scientist, working without knowledge of his colleague's labours, also wishes to test $t \,\&\, a$, but chooses a different experiment for this purpose, an experiment with only two possible outcomes: either e or $\sim e$. Of course, he obtains the latter, and having done so, must revise the refuted theory, to $t \,\&\, a'$, say. This scientist now notices that e' follows from the new theory and performs the orthodox experiment to verify the prediction. The new theory can now count a successful prediction to its credit, so it is not ad hoc.

But this is strange. We have arrived at opposite valuations of the very same theory on the basis of the very same observations, breaching at the same time what we previously called the Equivalence Condition and showing that the standard adhocness criteria are inconsistent. Whatever steps might be taken to resolve the inconsistency, it seems to us that one element ought to be removed, namely, the significance that the criteria attach to the order in which the theory and the evidence were thought up by a particular scientist, for this introduces into the principles of theory evaluation considerations concerning the state of scientists' minds that are irrelevant and incongruous in a methodology with pretensions to objectivity. No such considerations enter the corresponding Bayesian evaluations.

The Bayesian approach, incidentally, explains why people often react with instant incredulity, even derision, when certain ad hoc hypotheses are advanced. Is it likely that their amusement comes from perceiving, or even thinking they perceive, that the hypotheses lead to no new predictions? Surely they are simply struck by the utter implausibility of the claims.

Independent Evidence

The adhocness criteria are formulated in terms that refer to 'independent' evidence, yet this notion is always left vague and intuitive. How can it be made more precise? Probabilistic independence cannot fit the case. For suppose theory h was advanced in response to a refutation by e' and that h both explains that evidence and makes the novel prediction e''. It is the general opinion, certainly shared by Popperians, and also a consequence of Bayes's theorem, that e'' confirms h, provided it is sufficiently

improbable, relative to already available information. As discussed earlier in this chapter, such confirmation occurs, in particular, when $P(e'' \mid h \,\&\, e') > P(e'' \mid e')$. But this inequality can hold without e'' and e' being independent in the probabilistic sense.

Logical independence is also not the point here, for e'' might be independent from e' in this sense through some trivial difference, say, by relating to a slightly different place or moment of time. And in that case, e'' would not necessarily confirm or add credibility to h. For, as is intuitive, new evidence supports a theory significantly only when it is significantly different from known results, not just trivially different in the logical sense described. It is this intuition that appears to underlie the idea of independence used in the adhocness criteria.

That 'different' or 'varied' evidence supports a hypothesis more than a similar volume of homogeneous evidence is an old and widely held idea. As Hempel (1966, p. 34) put it: "the confirmation of a hypothesis depends not only on the quantity of the favourable evidence available, but also on its variety: the greater the variety, the stronger the resulting support". So, for example, a report that a stone fell to the ground from a certain height in such-and-such time on a Tuesday is similar to that relating to the stone's fall on a Friday; it is very different, however, from evidence of a planet's trajectory or of a fluid's rise in a particular capillary tube. But although it is often easy enough to classify particular bodies of evidence as either similar or varied, it is not easy to give the notions a precise analysis, except, in our view, in probabilistic terms, in the context of Bayesian induction.

The similar instances in the above list are such that when one of them is known, any other would be expected with considerable confidence. 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 is superfluous to cite more than a small, representative sample of such observations in evidence (see Urbach 1987, pp. 160–64). We are not concerned to give an exhaustive analysis of the intuitive notion, which is probably too vague for that to be possible, but are interested in that aspect of evidential similarity that is pertinent to confirmation. Bacon's

observations seem to capture this aspect and we may interpret his idea in probabilistic terms by saying that if two items of evidence, e_2 and e_1, are similar, then $P(e_2 \mid e_1) \approx 1$; when this condition holds, e_2 provides little support for any hypothesis if e_1 has already been cited as evidence. When the pieces of evidence are dissimilar, then $P(e_2 \mid e_1)$ is significantly less than 1, so that e_2 now does add a useful amount of confirmation to any already supplied by e_1. Clearly this characterization allows for similarity to be analysed in terms of degree.

To summarize, the non-Bayesian way of appraising hypotheses, and thereby of solving the Duhem problem, through the notion of adhocness is ungrounded in epistemology, has highly counter-intuitive consequences, and relies on a concept of independence amongst items of evidence that seems unsusceptible to analysis, except in Bayesian terms. In brief, it is not a success.

4.h | Designing Experiments

Why should anyone go to the trouble and expense of performing a new experiment and of seeking new evidence? The question has been debated recently. For example, Maher (1990) argues that since evidence can neither conclusively verify nor conclusively refute a theory, Popper's scientific aims cannot be served by gathering fresh data. And since a large part of scientific activity is devoted to that end, if Maher is right, this would constitute yet another serious criticism of Popper's philosophy. Of more concern to us is Miller's claim (1991, p. 2) that Bayesian philosophy comes up against the same difficulty:

> If e is the agent's total evidence, then $P(h \mid e)$ is the value of his probability and that is that. What incentive does he have to change it, for example by obtaining more evidence than he has already? He might do so, enabling his total evidence to advance from e to e^{-}; but in no clear way would $P(h \mid e^{-})$ be a better evaluation of probability than $P(h \mid e)$ was.

But the purpose of a scientific investigation, in the Bayesian view, is not to better evaluate inductive probabilities. It is to diminish uncertainty about a certain aspect of the world.

Suppose the question of interest concerns some parameter. You might start out fairly uncertain about its value, in the sense that your probability distribution over its range of possible values is fairly diffuse. A suitable experiment, if successful, would furnish evidence to lessen that uncertainty by changing the probability distribution, via Bayes's theorem, making it now more concentrated in a particular region; the greater the concentration and the smaller the region the better. This criterion has been given a precise expression by Lindley (1956), in terms of Shannon's characterization of information, and is discussed further in Howson 2002. Lindley showed that in the case where knowledge of a parameter θ is sought, provided the density of x varies with θ, any experiment in which x is measured has an expected yield in information. But, of course, this result is compatible with a well-designed experiment (with a high expected information yield) being disappointingly uninformative in a particular case; and by the same token, a poor experiment may be surprisingly productive of information.

In deciding whether to perform a particular experiment, at least three other factors should be taken into account: the cost of the experiment; the morality of performing it; and the value, both theoretical and practical, of the hypotheses one is interested in. Bayes's theorem, of course, cannot help here.

4.i | Under-Determination and Prior Probabilities

We pointed out in Chapter 1 that any data are explicable by infinitely many, mutually incompatible theories, a situation that some philosophers have called the 'under-determination' of theories by data. For example, Galileo carried out numerous experiments on freely falling bodies, in which he examined how long they took to descend various distances. His results led him to propound the well-known law: $s = a + ut + \frac{1}{2}gt^2$, where s is the distance fallen by the body in time t, and a, u and g are constants. Jeffreys (1961, p. 3) pointed out that without contradicting his own experimental results, Galileo might instead have advanced as his law:

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

where $t_1, t_2, \ldots, t_n$ are the elapsed times of fall that Galileo recorded in each of his experiments; a, u and g have the same values as above; and f is any function that is not infinite at any of the values $t_1, t_2, \ldots, t_n$. Jeffreys's modification therefore represents an infinity of alternatives to the orthodox theory, all implying Galileo's data, all mutually contradictory, and all making different predictions about future experiments.

There is a similar example due to Goodman (1954; for a lively and illuminating discussion, see Jeffrey 1983, pp. 187–190). He noted that the evidence of many green emeralds, under varied circumstances, would suggest to most observers that all emeralds are green; but he pointed out that that hypothesis bears the same relation to the evidence as does a type of hypothesis that he formulated as 'All emeralds are grue'. Goodman defined something as 'grue' when it was either observed before the present time ($T = 0$) and was green, or was not observed before that time and was blue. Clearly there are infinitely many grue-type predicates and infinitely many corresponding hypotheses, each associated with a different value of $T > 0$. All the current evidence of green emeralds is implied by both the green-hypothesis and the grue variants, yet not more than one of the hypotheses could be true.

As Jeffreys put it, there is always "an infinite number of rules that have held in all previous cases and cannot possibly all hold in future ones." This is a problem for those non-Bayesian scientific methods that regard a theory's scientific value as determined just by $P(e \mid h)$ and, in some versions, by $P(e)$. Such philosophical approaches, of which Popper's is one example, and maximum-likelihood estimation (Chapter 7, section **e**) another, would have to regard the standard law of free fall and Jeffreys's peculiar alternatives as equally good scientific theories relative to the evidence that was available to Galileo, and similarly with Goodman's strange hypotheses concerning emeralds, although these are judgments with which no scientist would agree.

In the Bayesian scheme, if two theories explain the evidence equally well, in the sense that $P(e \mid h_1) = P(e \mid h_2)$, this simply means that their posterior probabilities are in the same ratio as their priors. So theories, such as the contrived variants of Galileo's law and the Goodman grue-alternatives, which have the same

relation to the evidence as the orthodox theories and yet are received with incredulity, must have much lower prior probabilities. The role of prior probabilities also accounts for the important feature of scientific reasoning that scientists often prefer a theory that explains the data imperfectly, in the sense that $P(e \mid h) < 1$, to an alternative that explains them perfectly. This occurs when the better explanatory power of the alternative is offset by its inferior prior probability (Jeffreys 1961, p. 4).

This Bayesian account is of course only partial, for we can provide no general account of the genesis of prior probabilities. In some situations, the prior may simply be the posterior probability derived from earlier results and an earlier prior. Sometimes, when there are no such results, a prior probability may be created through what we know from other sources. Consider, for instance, a theory that makes some assertion about a succession of events in the development of a human society; it might, for example, say that the elasticity of demand for herring is constant over a particular period, or that the surnames of all future British prime ministers and American presidents will start with the letter *B*. These theories could possibly be true, but are immensely unlikely to be so. And the reason for this is that the events they describe are the causal effects of numerous, independent processes, whose separate outcomes are improbable. The probability that all the processes will turn out to favour one of the theories in question is therefore the product of many small probabilities and so is itself very small indeed (Urbach 1987b). But the question of how the probabilities of the causal factors are estimated remains. 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 form of reasoning. For instance, the 'simplicity' of a hypothesis has been thought to have an influence on its initial probability. This and other possibilities are discussed in Chapter 9.

4.j | Conclusion

The various, mostly familiar aspects of scientific reasoning that we have examined have all shown themselves to correspond nat-

urally to aspects of Bayesian logic, whereas non-Bayesian accounts fail more or less completely. So far, we have concentrated chiefly on deterministic theories. We shall see in the next and following chapters that the Bayesian approach applies equally well to statistical reasoning.

CHAPTER 5
Classical Inference: Significance Tests and Estimation

In the last chapter, we showed how leading aspects of scientific reasoning are illuminated by reference to Bayes's theorem, confining our attention, however, mainly to deterministic theories. We now consider theories that are not deterministic but probabilistic, or statistical. From the Bayesian viewpoint the division is artificial and unnecessary, the two cases differing only in regard to the probability of the evidence relative to the theory, that is, $P(e \mid h)$, which figures in the central theorem: when h is deterministic, this probability is either 1 or 0, depending on whether h entails e or is refuted by it; when h is statistical, $P(e \mid h)$ typically takes an intermediate value. The uniform treatment that this affords is unavailable in non-Bayesian methodologies, whose advocates have instead developed a specific system, known as *Classical Statistical Inference* or sometimes as *Frequentism*, to deal with statistical theories.

This system, with its 'significance tests', 'confidence intervals', and the rest, swept the board for most of the twentieth century, and its influence is still considerable. The challenge to Bayesian methodology posed by Frequentism requires an answer, and this we shall give in the present and succeeding chapters.

5.a Falsificationism in Statistics

The simple and objective mechanism by which a hypothesis may, under certain circumstances, be logically refuted by observational evidence could never work with statistical hypotheses, for these ascribe probabilities to possible events and do not say of any that they will or will not actually occur. The fact that statistical theories have a respected place in science and are regularly tested and

evaluated through experiment is therefore an embarrassment to the methodology of falsificationism. In consequence, defenders of that methodology have tried to take account of statistical theories by modifying its central dogma.

The modified idea acknowledges that a statistical hypothesis is not strictly falsifiable, and what it proposes is that when an event occurs to which the hypothesis attaches a sufficiently small probability, it should be *deemed* false; scientists, Popper said, should make "a methodological decision to regard highly improbable events as ruled out—as prohibited" and he talked of hypotheses then being "practically falsified" (1959a, p. 191). The mathematician and economist Cournot (1843, p. 155) expressed the same idea when he said that events of sufficient improbability "are rightly regarded as physically impossible".

But is it right? After all, a distinctive feature of statistical hypotheses is that they do not rule out events that they class as improbable. For example, the Kinetic Theory attaches a tiny probability to the event of ice spontaneously forming in a hot tub of water, but does not rule it out; indeed the fact that the theory reveals so strange an event as a possibility, contrary to previous opinion, is one of its especially interesting features. And even though this particular unlikely event may never materialize, immensely improbable events, which no one would regard as refuting the Kinetic Theory, do occur all the time, for instance, the spatial distribution at a particular moment of the molecules in this jug of water.

Or take the simple statistical theory that we shall frequently use for the purpose of illustration, which claims of some particular coin that it has a physical probability, constant from throw to throw, of $\frac{1}{2}$ of landing heads and the same probability of landing tails (the coin is said then to be 'fair'). The probability of any particular sequence of heads and tails in, say, 10,000 tosses of the coin is $2^{-10,000}$, a minuscule value, yet it is the probability of every possible outcome of the experiment, one of which will definitely occur. The implication of the Cournot-Popper view that this definite occurrence should be regarded as physically impossible is clearly untenable.

5.b Fisherian Significance Tests

Fisher was inspired by both the falsificationist outlook and the ideal of objectivity when, building on the work of Karl Pearson and W.S. Gossett (the latter, writing under the pen name 'Student'), he developed his system of *significance tests* for testing statistical theories. Fisher did not postulate a minimal probability to represent physical impossibility, and so avoided the problem that destroys the Cournot-Popper approach. His proposal, roughly speaking, was that a statistical hypothesis should be rejected by experimental evidence when it is, on the assumption of that hypothesis, contained in a certain set of outcomes that are *relatively* unlikely, relative, that is, to other possible outcomes of the experiment.

Before assessing how well they are suited to their task, let us set out more precisely the nature of Fisher's significance tests, which we shall illustrate using, as the hypothesis under test (what Fisher called the *null hypothesis*), the fair-coin hypothesis mentioned above. To perform the test, an experiment must be devised: in our example, it will involve flipping the coin a predetermined number of times, say 20, and noting the result; this result is then analysed in the following four stages.

1 First, specify the *outcome space*, that is, all the results that the experiment could have produced. In our example, this would normally be taken to comprise the 2^{20} possible sequences of 20 heads or tails. (We examine the assumptions underlying the specification of any outcome space in the next section when we discuss 'stopping rules'.) The result of a coin-tossing experiment would not normally be reported as a point in the outcome space just described but would be summarized in some numerical form, and for the purpose of our example, we shall select r, the number of heads in the outcome. Such a numerical summary when used in a significance test is known as a *test-statistic*; it is formally a random variable, as defined in Chapter 2. (We shall presently discuss the basis upon which test-statistics are chosen.)

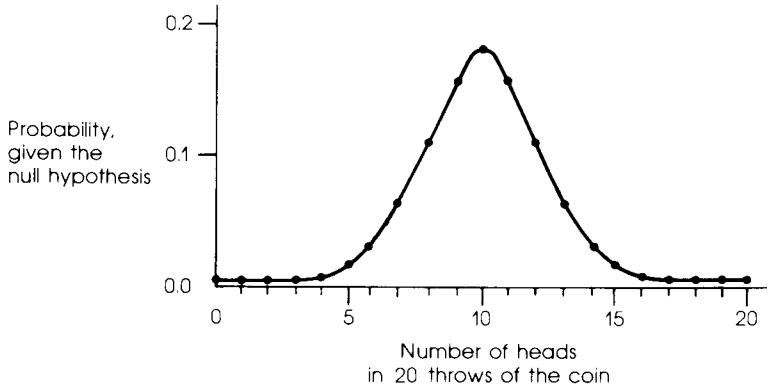
2 Next, calculate the probability, relative to the null hypothesis, of each possible value of the test-statistic—its *sampling distribution*. In general, if the probability of getting a head in a

coin-tossing experiment is p, and of getting a tail is q, then r heads will appear in n tosses of the coin with probability ${}^nC_r p^r q^{n-r}$.[1] In the present case, $p = q = \frac{1}{2}$ and $n = 20$. The required probabilities can now be directly calculated; they are shown in Table 5.1 and also displayed graphically below.

TABLE 5.1

The Probabilities of Obtaining *r* Heads in a Trial consisting of 20 Tosses of a Fair Coin

Number of Heads (r)	*Probability*	*Number of Heads* (r)	*Probability*
0	9×10^{-7}	11	0.1602
1	1.9×10^{-5}	12	0.1201
2	2×10^{-4}	13	0.0739
3	0.0011	14	0.0370
4	0.0046	15	0.0148
5	0.0148	16	0.0046
6	0.0370	17	0.0011
7	0.0739	18	2×10^{-4}
8	0.1201	19	1.9×10^{-5}
9	0.1602	20	9×10^{-7}
10	0.1762		



[1] This familiar fact is demonstrated in standard statistics textbooks. ${}^nC_r$ is equal to $\frac{(n!)}{(n-r)!\,r!}$.

3 The third stage of Fisher's analysis requires us to look at all the results which *could have* occurred and which, relative to the null hypothesis, are, as Fisher put it, "more extreme than" the result that did occur. In practice, this vague expression is interpreted probabilistically, the requirement then being that we examine possible outcomes of the trial which, relative to the null hypothesis, have a probability less than or equal to the probability of the actual outcome. We should then calculate the probability (p^*) that the experimental result will fall within this group. (p^* is often called the *p-value* of the result.)

To illustrate, suppose our experiment produced 4 heads and 16 tails, which we see from the table occurs, if the null hypothesis is true, with probability 0.0046. The results with less or equal probability to this are $r = 4, 3, 2, 1, 0$ and $r = 16, 17, 18, 19, 20$ and the probability of any one of them occurring is the sum of their separate probabilities, *viz*:

$$p^* = 2 \times (0.0046 + 0.0011 + 2 \times 10^{-4} + 1.9 \times 10^{-5} + 9 \times 10^{-7}) = 0.012.$$

4 A convention has grown up, following Fisher, to *reject* the null hypothesis just in case $p^* \leq 0.05$. However, some statisticians recommend 0.01 or even 0.001 as the critical probability. The critical probability that is adopted is called the *significance level* of the test and is usually labelled α. If an experimental result is such that $p^* \leq \alpha$, it is said to be *significant at the α significance level*, and the null hypothesis is said to be *rejected at the α (or 100α percent) level.*

In our example, the coin produced 4 heads when flipped 20 times, corresponding to $p^* = 0.012$; since this is below 0.05, the null hypothesis should be rejected at the 0.05 or 5 percent level. But a result of 6 heads and 14 tails, with $p^* = 0.115$, would not be significant, and so the null hypothesis should then not be rejected at that level.

This simple example illustrates the bare bones of Fisher's approach. It is, however, not always so easy to apply in practice. Take the task often treated in statistics textbooks of testing whether two populations have the same means, for instance,

whether two groups of children have the same mean IQ. It may not be feasible to take measurements from every child, in which case, the recommended procedure is to select children at random from each of the groups and compare their IQs. But to perform a significance test on the results of this sampling one needs a test-statistic with a determinate and known distribution and these are often difficult to find. A solution was found in the present case by 'Student', who showed that provided the experimental samples were sufficiently large to ensure approximate normality, the so-called *t*-statistic[2] has the appropriate properties for use in a significance test.

Which Test-Statistic?

Fisher's theory as so far expounded is apparently logically inconsistent. This is because different random variables may be defined on any given outcome space, not all of them leading to the same conclusion when used as the test-statistic in a significance test; one test-statistic may instruct you to reject some hypothesis when another tells you not to.

We can illustrate this very simply in relation to our coin-tossing experiment. We there chose the number of heads in the outcome as the test-statistic, which, with 20 throws of the coin, takes values from 0 to 20. Now define a new statistic, r', with values from 0 to 18, derived from the earlier statistic by grouping the results as indicated in Table 5.2. In this slight modification, the outcome *5 heads* and the outcome *10 heads* are counted as a single result whose probability is that of obtaining either one of these; similarly, for the results *14* and *15 heads*. This new statistic is artificial, having no natural meaning or appeal, but according to the definition, it is a perfectly proper test-statistic.

It will be recalled that previously, with the number of heads as the test-statistic, the result *6 heads, 14 tails* was *not* significant at the 0.05 level. It is easy to see that using the modified statistic, this result now *is* significant at that level ($p^* = 0.049$). Hence Fisher's principles as so far described tell us both to reject and not to reject the null hypothesis, which is surely impossible. Clearly

[2] See Section 6.c for more details of the *t*-statistic.

TABLE 5.2

The Probability Distribution of the r'-Statistic

Value of Statistic (r')	*Probability*	*Value of Statistic (r')*	*Probability*
0 (0 heads)	9×10^{-7}	10 (11 heads)	0.1602
1 (1 heads)	1.9×10^{-5}	11 (12 heads)	0.1201
2 (2 heads)	2×10^{-4}	12 (13 heads)	0.0739
3 (3 heads)	0.0011	13 (14 or 15 heads)	0.0518
4 (4 heads)	0.0046	14 (16 heads)	0.0046
5 (6 heads)	0.0370	15 (17 heads)	0.0011
6 (7 heads)	0.0739	16 (18 heads)	2×10^{-4}
7 (8 heads)	0.1201	17 (19 heads)	1.9×10^{-5}
8 (9 heads)	0.1602	18 (20 heads)	9×10^{-7}
9 (5 or 10 heads)	0.1910		

test-statistics need some restriction that will ensure that all permissible ones lead to similar conclusions. And any such restriction must be recommended by more than the consistency it brings; it must produce the right consistent result, if there is one, for the right reasons, if there are any.

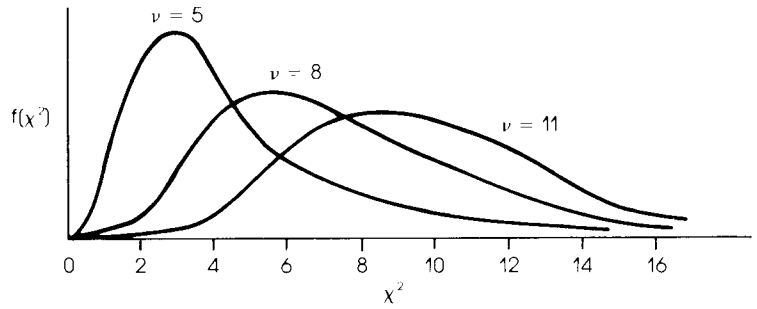
The Chi-Square Test

A striking illustration of the difficulties posed by the multiplicity of possible test-statistics is the chi-square (or χ^2) goodness-of-fit test, which bulks large in the literature and is widely used to test hypotheses that ascribe probabilities to several different types of event, for example, to the outcomes of rolling a particular die. Suppose the die were rolled n times and landed with a six, five, etc. showing uppermost with frequencies $O_6, O_5, \ldots, O_1$. If p_i is the probability that the null hypothesis ascribes to the outcome i, then np_i is the *expected frequency* (E_i) of that outcome. The null hypothesis is tested by the following so-called *chi-square statistic*:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i},$$

the sum being taken over all the possible outcomes of the trial.

Karl Pearson discovered the remarkable fact, very helpful for its application to significance tests, that the probability distribution of this statistic is practically independent of the unknown probabilities and of n and is dependent just on the number, ν, of the test's so-called *degrees of freedom*, where $\nu = J - 1$, and J is the number of separate cells into which the outcome was divided when calculating χ^2. The probability density distributions of χ^2, for various values of ν, are roughly as follows:



We may illustrate the χ^2-test with a simple example. Let the null hypothesis assert that a particular die is 'true', that is, has equal probabilities, of $\frac{1}{6}$, constant from throw to throw, of falling with each of its sides uppermost. Now consider an experiment involving 600 rolls of the die, giving the results, say: *six* (90), *five* (91), *four* (125), *three* (85), *two* (116), *one* (93).

To perform a chi-square test, we must calculate χ^2 for these data, as follows ($E_i = 600 \times \frac{1}{6} = 100$, for each i):

$$\chi^2 = \frac{1}{100}[(100-90)^2 + (100-91)^2 + (100-125)^2 + (100-85)^2 + (100-116)^2 + (100-93)^2] = 13.36.$$

Since the outcome involves six cells, the number of degrees of freedom is five and, as can be roughly gauged from the above sketches and more precisely established by consulting the appropriate tables, the probability of obtaining a value of χ^2 as large or larger than 13.36 is less than 0.05, so the result is significant, and the null hypothesis must therefore be rejected at the corresponding significance level.

Chi-square tests are also used to test theories asserting that some population has a particular, continuous probability distribution, such as the normal distribution. To test such a theory, the range of possible results of some sampling trial would be divided into several intervals and the numbers of subjects falling into each would be compared with the 'expected' number, proceeding then as with the example of the die.

Although the test has been much further developed and with great technical ingenuity, it is, we believe, vitiated by the absence of any principled rule for partitioning the outcomes into separate intervals or cells, for not all partitions lead to the same inferences when the significance test is applied. For instance, in our die-rolling example, if we had based χ^2 on just three cells, formed, say, by combining the pairs of outcomes [*six*, *five*], [*four*, *three*], and [*two*, *one*], the result would not now be significant at the 5 percent level.

This problem is rarely taken up in expositions of the chi-square test. When it is, it is resolved by considerations of convenience, not epistemology. For instance, Kendall and Stuart (1979, p. 457) argued that the class boundaries should be drawn so that each cell has the same probability (relative to the null hypothesis) of containing the experimental outcome, and they defended this rule on the epistemically irrelevant grounds that it is "perfectly definite and unique". But it is not even true that the equal-probability rule leads to a unique result, as we can see from our last example. We there considered partitioning the outcomes of the die-rolling experiment into three pairs: there are in fact fifteen distinct ways of doing this, all satisfying the equal-probability rule, and only two of them render the results significant at the 5 percent level.

The complacency of statisticians in the face of this difficulty is remarkable. Although Hays and Winkler (1970, p. 195) warn readers repeatedly and emphatically that "*the arrangement into population class intervals is arbitrary*", their exposition proceeds without recognizing that this renders the conclusions

of a chi-square test *equally* arbitrary. Cochran (1952, p. 335) claimed that the problem is a "minor" one, which merely calls for "more standardization in the application of the test". But standardization would only institute the universal application of arbitrary principles and would not address the central problem of the chi-square test, which is how to set it on a firm epistemic basis. No such basis appears to exist, and in view of this, the test should be abandoned.

It might be argued that, despite its epistemic difficulties, there are strong indications that the chi-square test is sound, because the conclusions that are in practice drawn from it generally fit so well with intuition. In many cases, intuition is indeed satisfied but the test can also produce quite counter-intuitive inferences, and once one such inference has been seen, it is easy to generate more. Suppose, for instance, that the above trial with the die had given the results: *six* (123), *five* (100), *four* (100), *three* (100), *two* (100), *one* (77). In a test of the hypothesis that the die was true, χ^2 takes the value 10.58, which is not significant at the 5 percent level, and any statistician who adopted this as the critical value would not be obliged to reject the null hypothesis, even though the results of the trial tell us pretty clearly that it is quite wrong.[3]

In the examples we have so far considered, only large values of χ^2 were taken as grounds for rejecting a hypothesis. But for Fisher both extremities of any test-statistic's distribution were critical. In his view, a null hypothesis is "as definitely disproved" when the observed and the expected frequencies are very similar, leading to a very small χ^2, as it is when the frequencies are sharply discrepant and χ^2 is large (Fisher 1970, Section 20). This forms the basis of Fisher's famous criticism of Mendel's experimental results, which we discussed above in 4.e. Those results, he said, were "too good to be true", that is to say, although they seemed to be in close accord with Mendelian theory, and were usually taken to be so, they corresponded to χ^2 values that were sufficiently small to imply its rejection in a significance test. For Fisher the chi-square test had to override intuitions in this case. But this is not the universal opinion amongst classical statisti-

[3] Good 1981, p.161, makes this point.

cians. For example, Stuart (1954) maintained that a small χ^2 is critical only if all "irregular" alternatives to the null hypothesis have been ruled out, where the irregularity might involve "variations due to the observer himself", such as "all voluntary and involuntary forms of falsification". Indeed, the Fisherian idea that a null hypothesis can be tested in isolation, without considering rival hypotheses, is not now widely shared and the predominant form of the significance test, that of Neyman and Pearson, which we discuss shortly, requires hypotheses to be tested against, or in the context of, alternative hypotheses.

Sufficient Statistics

It is sometimes claimed that consistency may be satisfactorily restored to Fisher's significance tests by restricting test-statistics to so-called *minimal-sufficient statistics*, because of their standard interpretation as containing all the information that is relevant to the null hypothesis and none that is irrelevant. We shall argue, however, that this interpretation is unavailable to Fisher, that there are no grounds for excluding irrelevant information from a test, and that the difficulty confronting Fisherian principles is unconnected with the amount of information in the test-statistic, but lies elsewhere.

Let us first examine the concept of a sufficient statistic. Some statistics clearly abstract more information from the outcomes than others. For instance, tossing a coin four times will result in one of the sixteen sequences of heads and tails (*HHHH*), (*THHH*), . . . , (*TTTT*), and a statistic that assigns distinct numbers to each element of this outcome space preserves all the information produced by the experiment. But a statistic that records only the number of heads thereby discards information, so if you knew only that it took the value 3, say, you could not determine from which of the four different outcomes containing 3 heads it was derived. Whether some of the discarded information is relevant to an inference is a question addressed by the theory of sufficiency.

A sample statistic, t, is said to be *sufficient*, relative to a parameter of interest, θ, if the probability of any particular member of the outcome space, given t, is independent of θ. In our

example, the statistic representing the number of heads in the outcome is in fact sufficient for θ, the physical probability of the coin to land heads, as can be simply shown. The outcome space of the coin-tossing experiment consists of sequences $x = x_1, \ldots, x_n$, where each x_i denotes the outcome either *heads* or *tails*, and $P(x \mid t)$ is given as follows, remembering that, since the value of t is logically implied by x, $P(x \,\&\, t) = P(x)$:

$$P(x \mid t) = \frac{P(x \,\&\, t)}{P(t)} = \frac{P(x)}{P(t)} = \frac{\theta^r(1-\theta)^{n-r}}{{}^nC_r\theta^r(1-\theta)^{n-r}} = \frac{1}{{}^nC_r}$$

Since the binomial term, ${}^nC_r$, is independent of θ, so is $P(x \mid t)$; hence, t is sufficient for θ.

It seems natural to say that if $P(x \mid t)$ is the same whatever the parameter value, then x "can give us no information about θ that the sufficient statistic has not already given us" (Mood and Graybill 1963, p.168). Certainly Fisher (1922, p. 316) understood sufficiency that way: "The Criterion of Sufficiency", he wrote, is the rule that "the statistic chosen should summarize the whole of the relevant information supplied by the sample". But natural as it seems, this interpretation is unavailable to Fisher, for a hypothesis subjected to one of his significance tests may be rejected by one sufficient statistic and not by another. Our coin-tossing example illustrates this, for the statistic that summarizes the outcome as the number of heads in the sample, and the statistic that assigns separate numbers to each member of the outcome space are both sufficient, as is the artificial statistic r', described above, though these statistics do not generally yield the same conclusion when used in a Fisherian test of significance.

Since the sufficiency condition does not ensure a unique conclusion, the further restriction is sometimes argued for (for example by Seidenfeld 1979, p. 83) that the test-statistic should be *minimal-sufficient*; that is, it should be such that any further reduction in its content would destroy its sufficiency. A minimal-sufficient statistic is thought of as containing all the information supplied by the sample that is relevant, and none that is irrelevant. But this second restriction has received no adequate defence; indeed, it would be surprising if a case could be made for it, for if information is irrelevant, it should make no difference to a test, so there should be no need to exclude it. It is curious that, despite the

almost universal lip service paid to the sufficiency condition, the principal statistics that are in practice used in significance tests—the χ^2, t and F statistics—are none of them sufficient, let alone minimal-sufficient (Pratt 1965, pp. 169–170).

The idea of restricting admissible statistics according to their information content seems in any case misconceived as a way of saving Fisherian significance tests. For Neyman (1952, pp. 45–46) has shown that where the null hypothesis describes a continuous probability density distribution over the space, there may be pairs of statistics that are related by a 1-1 transformation, such that only one of them leads to the rejection (at a specified significance level) of the null hypothesis. Since these statistics necessarily carry the same information, there must be some other source of the trouble.

5.c | Neyman-Pearson Significance Tests

Fisher's significance tests were designed to provide for the statistical case something akin to the falsification available in the deterministic case; hence his insistence that the tests should operate on isolated hypotheses. But as we indicated earlier, statistical hypotheses cannot be refuted and, as we show later (Section 5.d), Fisher's own analysis of and arguments for a quasi-refutation are quite unsatisfactory. For this reason, Neyman felt that a different epistemic basis was required for statistical tests, in particular, one that introduces rival hypotheses into the testing process. The version of significance tests that he and Pearson developed resembled Fisher's however, in according no role to prior or posterior probabilities of theories, for they were similarly opposed to Bayesian methodology.

In setting out the Neyman-Pearson method, we shall first consider the simplest cases, where only two hypotheses, h_1 and h_2, are in competition. Neyman-Pearson tests permit two kinds of inference: either a hypothesis is rejected or it is accepted. And such inferences are subject to two sorts of error: you could regard h_1 as false when in fact it is true, or accept h_1 (and, hence, reject h_2) when it is false. When these errors can be distinguished by their gravity, the more serious is called a *type I* error

and the less serious a *type II* error. The seriousness of the two types of error is judged by the practical consequences of acting on the assumption that the rejected hypothesis is false and the accepted one true. For example, suppose two alternative hypotheses concerning a food additive were admitted, one that the substance is safe, the other that it is highly toxic. Under a variety of circumstances, it would be less dangerous to assume that a safe additive was toxic than that a toxic one was safe. Neyman and Pearson, adapting Fisher's terminology, called the hypothesis whose mistaken rejection is the more serious error the *null hypothesis*, and where the errors seem equally serious, either hypothesis may be so designated.

The possibilities for error are summed up in Table 5.3.

TABLE 5.3

Decision	*True Hypothesis*	
	h_1	h_2
Reject h_1	Error	/
Accept h_1	/	Error

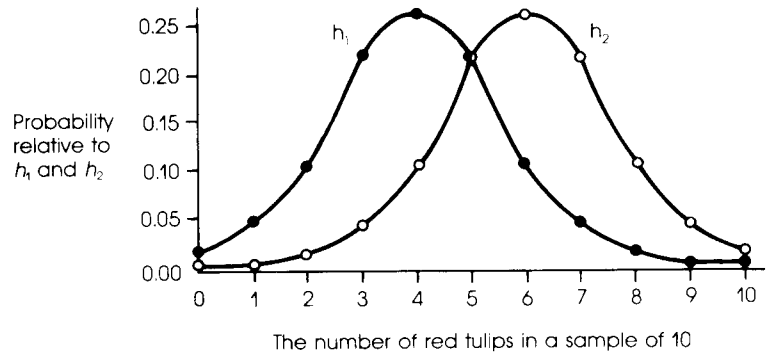
The Neyman-Peason approach aims to minimize the chance of committing both types of error. We will examine the Neyman-Pearson approach through an example borrowed from Kyburg (1974, pp. 26–35). The label on a particular consignment of tulip bulbs has been lost and the purchaser cannot remember whether it was the one that contained 40 percent of the red- and 60 percent of the yellow-flowering sort, or 40 percent of the yellow and 60 percent of the red. We shall designate these possibilities h_1 and h_2, respectively, and treat the former as the null hypothesis. An experiment to test these hypotheses might involve planting a predetermined number of bulbs, say 10, that have been randomly selected from the consignment, and observing which grow red and which yellow. The testing procedure is similar to Fisher's and involves the following steps.

First, specify the outcome space, which in the present case may be considered to comprise 2^{10} sequences, each sequence indicating the flower-colour of the tulip bulb that might be selected first, second, and so on, down to the tenth. Next, a test-statistic that summarizes the outcome in numerical form needs to be stipulated, and in this example we shall take the number of reds appearing in the sample for this purpose. (We discuss the basis for these arbitrary-seeming stipulations below.) Thirdly, we must compute the probabilities of each possible value of the test-statistic, relative to each of the two rival hypotheses. If, as we shall assume, the consignment of tulip bulbs is large, the probability of selecting r red-flowering bulbs in a random sample of n is approximated by the familiar binomial function ${}^nC_r p^r q^{n-r}$. We are assuming here that the probability, p, of selecting a red-flowering bulb is constant, an assumption that is more approximately true, the larger the population of bulbs. In the present case, h_1 corresponds to $p = 0.40$, and h_2 to $p = 0.60$. The sampling distributions for the imagined trial, relative to the two hypotheses, are given in Table 5.4 and displayed graphically, below.

TABLE 5.4

The Probabilities of Selecting *r* Red- and (10 – *r*) Yellow-flowering Tulips

Outcome (Red, Yellow)	h_1 $(p = 0.40)$	h_2 $(p = 0.60)$
0, 10	0.0060	0.0001
1, 9	0.0403	0.0016
2, 8	0.1209	0.0106
3, 7	0.2150	0.0425
4, 6	0.2508	0.1115
5, 5	0.2006	0.2006
6,4	0.1115	0.2508
7, 3	0.0425	0.2150
8, 2	0.0106	0.1209
9, 1	0.0016	0.0403
10, 0	0.0001	0.0060



Finally, the Neyman-Pearson method calls for a rule that will determine when to reject the null hypothesis. Consider the possibility of rejecting the hypothesis just in case 6 or more red-flowering plants appear in the sample. Then, if h_1 is true, the probability of a rejection may be seen from the table to be: 0.1115 + 0.0425 + 0.0106 + 0.0016 + 0.0001 = 0.1663, and this is therefore the probability of a type I error associated with the postulated rejection rule. This probability is called, as before, the significance level of the test, or its *size*. The probability of a type II error is that of accepting h_1 when it is false; on our assumption that one of the two hypotheses is true, this is identical to the probability of rejecting h_2 when it is true, which may be calculated from the table as 0.3664.

The *power* of a test is defined as 1 – *P*(type II error) and is regarded by advocates of this approach as a measure of how far the test 'discriminates' between the two hypotheses. It is also the probability of rejecting the null hypothesis when it is false, and in the present case has the value 0.6336.

In selecting the size and power of any test, a natural ideal might seem to be to try to minimize the former and maximize the latter, in order to reduce as far as possible the chances of both types of error. We shall, in due course, consider whether an ideal couched in terms of type I and type II errors is suited to inductive reasoning. We have to note straightaway, though, that the ideal is incoherent as it stands, for its twin aims are incompatible: in most cases, a diminution in size brings with it a contraction in power, and vice versa. Thus, in our example, if the rejection rule were

changed and h_1 rejected in the event of at least 7 red tulips in the sample, the size of the test would be reduced from 0.1663 to 0.0548, but its power would also be lower – 0.3823, compared with 0.6336. We see then that while the revised test has a smaller size, this advantage (as it is judged) is offset by its smaller power. For this reason, Neyman and Pearson proposed instead that one first fix the size of a test at an appropriate level and, thus constrained, then maximize its power.

Randomized Tests

It is generally held amongst classical statisticians that the size of a significance test should not exceed 0.05 and, for a reason we shall describe later, practitioners are often exhorted always to employ roughly the same significance levels. But with the methods introduced so far the size of a test cannot always be chosen at will. For this purpose, randomized tests have been devised, which Kyburg has lucidly explained in the context of the example we have been discussing. Suppose a test of size 0.10 were desired. Let the two tests considered above be labelled 1 and 2. As we showed, they have the following characteristics:

TABLE 5.5

	Probability of a type I error	*Probability of a type II error*	*Power*
Test 1	0.1663	0.3664	0.6336
Test 2	0.0548	0.6177	0.3823

Imagine, now, a third test which is carried out in the following manner: a pack of 200 cards, of which 119 are red and 81 black, is well shuffled, and one of these cards is then randomly selected. If the selected card is black, test 1 is applied, and if red, test 2. This *mixed* or *randomized test* has the required size of 0.10, given by

$$\frac{81}{200} \times 0.1663 + \frac{119}{200} \times 0.0548 = 0.100.$$

The corresponding probability of a type II error is similarly calculated to be 0.5159; so the power of the mixed test is 0.4841.

Readers might be surprised by the implication that inspecting a piece of coloured card, whose causal connexion to the tulip consignment is nil, can nevertheless provide an insight into its composition. Randomized tests are rarely if ever used, but they form a proper part of the Neyman-Pearson theory, so any criticism that they merit can quite correctly be re-directed to the Neyman-Pearson theory in general.

The Choice of Critical Region

An advantage that Neyman-Pearson significance tests enjoy over Fisher's is that they incorporate in a quite natural way a feature that Fisher seems to have adopted arbitrarily and in deference merely to apparent scientific practice, namely, to concentrate the critical region in (one or both of) the tails of the sampling distribution of outcomes. Fisher's reasoning seems to have been that evidence capable of rejecting a hypothesis must be very improbable and should lie in a region of very low probability (see Section 5.d). But Neyman pointed out that by this reasoning, Fisher could equally well have chosen for the rejection region a narrow band in the centre of a bell-shaped distribution as a broader band in its tails.

By contrast, in the Neyman-Pearson approach, the critical region is uniquely determined, according to a theorem known as the *Fundamental Lemma*. This states that the critical region of maximum power in a test of a null hypothesis, h_1, against a rival, h_2, is the set of points in the outcome space that satisfies the inequality:

$$\frac{P(x \mid h_1)}{P(x \mid h_2)} \leq k,$$

where k is a constant that depends on the hypotheses and on the significance level.[4] The probabilities may also be densities.

[4] Strictly speaking, the likelihoods $P(x \mid h_1)$, $P(x \mid h_2)$ should not be expressed here as conditional probabilities, for these presuppose that the hypotheses themselves have a probability, something that classical statisticians strenuously deny. Hence, they are sometimes written $P(x ; h)$ or $L(x \mid h)$. Bayesians, of course, need have no such qualms.

The lemma embraces randomized tests, the critical region then comprising those of the component non-randomized tests, which are selected at random, as already described.

Neyman-Pearson Tests and Sufficient Statistics

Neyman-Pearson tests have another fortunate consequence, namely, that for them sufficient statistics do contain all the relevant information. For if h_1 and h_2 ascribe different values to a parameter θ, and if t is a sufficient statistic relative to the outcomes $x = x_1, \ldots, x_n$, then, by definition, $P(x \mid t)$ is independent of θ, and it follows almost directly that

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

The above lemma tells us that the left-hand ratio does not exceed some number k; hence, the same holds also for the right-hand ratio. So if the outcome were summarized in terms of t, rather than x, the region of maximum power would comprise the same outcomes, and consequently, none of the information in x that is omitted from t is relevant to the significance test inference.

5.d Significance and Inductive Significance

'The null hypothesis was rejected at such-and-such a significance level' is a technical expression that simply records that an experimental result fell in a certain designated 'rejection region' of the outcome space. But what does it mean as an inductive conclusion about the hypothesis? There are three principal views on this amongst advocates of the significance test. None, we shall argue, is in the least satisfactory.

Fisher's View

Fisher took the process of logical refutation as the model for his significance tests. This is apparent in his frequently voiced claim that such tests could "disprove" a theory (for example, 1947, p. 16), and that "when used accurately, [they] are capable of rejecting or invalidating hypotheses, in so far as these are *contradicted by the data*" (1935; our italics).

Fisher seems to be saying here that statistical theories may actually be falsified, though, of course, he knew full well that this was impossible, and in his more careful accounts he took a different line.[5] The force of a test of significance, he said (1956, p. 39), "is logically that of the simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the theory of random distribution [i.e., the null hypothesis] is not true". But in thus avoiding an unreasonably strong interpretation, Fisher fell back on one that is unhelpfully weak, for the significant or critical results in a test of significance are by definition improbable, relative to the null hypothesis. Inevitably, therefore, a significant result is either a "rare chance" (an improbable event) or the null hypothesis is false, or both. And Fisher's claim amounts to no more than this empty truism.[6]

Significant Results and Decisions to Act

Neyman and Pearson proposed what is now the most widely adopted view, namely, that on 'accepting' a hypothesis after a significance test, one should act as if one believed it to be true, and if one 'rejects' it, one's actions should be guided by the assumption that it is false, "without", as Lindgren (1976, p. 306) put it, "necessarily being convinced one way or the other". Neyman and Pearson (1933, p. 142) defended their rule by saying that although

[5] The careless way that Fisher sometimes described the inductive meaning of a significant result is often encountered in statistics texts. For example, Bland (1987, p. 158) concludes from one such result "that the data are not consistent with the null hypothesis"; he then wanders to the further conclusion that the alternative hypothesis is "more likely".

[6] Hacking 1965, p.81, pointed this out.

a significance test "tells us *nothing* as to whether in a particular case h is true"[7], nevertheless

> it may often be proved that if we behave according to . . . [the rule], then in the long run we shall reject h when it is true not more, say, than once in a hundred times [when the significance level is 0.01], and in addition we may have evidence that we shall reject h sufficiently often when it is false [i.e., when the test's power is sufficiently large].

This is a surprising argument to encounter in this context. After all, the significance test idea was born out of the recognition that events with probability p cannot be proved to occur with any particular frequency, let alone with frequency p; indeed, they may never occur at all. This is acknowledged tacitly in the above argument, through the proviso "in the long run", a prevarication that suggests some largish but practically accessible number, yet at the same time also hints at the indefinite and infinite. The former suggestion is, as we have said, unsustainable; the latter would turn the argument into the unhelpful truism that with a significance level of 0.01, we would reject a true null hypothesis with probability 0.01. Either way, the argument does not uphold the Neyman-Pearson rejection rule.

There are also objections to the idea of acting as if a hypothesis were definitely true or definitely false when one is not convinced one way or the other. If, to go back to our earlier example, one were to reject the hypothesis that the tulip consignment contained 40 percent of the red variety and then act as if it were definitely false, there would be no incentive to repeat the experiment and every incentive to stake all one's worldly goods, and whatever other goods one might possess, on a wager offered at any odds on the hypothesis being true. The idea is clearly absurd.

Neyman (1941, p. 380), on the other hand, argued that there are in fact occasions when it is reasonable to behave as if what one believed to be false were actually true, and *vice versa*, citing the

[7] These are our italics. Neyman's view that no inductive inferences are licensed by sampling information is discussed in Section 5.f.2 below.

purchase of holiday insurance as a case in point. In making such a purchase, he said, "we surely act against our firm belief that there will be no accident; otherwise, we would probably stay at home". This, however, seems a perverse analysis of the typical decision to take out insurance. We surely do not firmly believe that there will be no accident when we go away, but regard the eventuality as more or less unlikely, depending on the nature of the holiday, its location, and so forth; and the degree of risk perceived is reflected in, for example, the sum we are prepared to lay out on the insurance premium.

Another example that is sometimes used to defend the idea of acting as if some uncertain hypothesis were true is industrial quality control.[8]

The argument is this. Suppose an industrialist would lose money by marketing a product-run that included more than a certain percentage of defective items. And suppose product-runs were successively sampled, with a view to testing whether they were of the loss-making type. In such cases, there could be no graduated response, it is claimed, since the product-run can either be marketed or not; but, the argument goes, the industrialist could be comforted by the thought that "in the long run" of repeatedly applying the same significance test and the same decision rule, only about, say, 5 percent of the batches marketed will be defective, and that may be a financially sustainable failure rate.

But this argument does not succeed, for the fact that only two actions are possible does not imply that only two beliefs can be entertained about the success of those actions. The industrialist might attach probabilities to the various hypotheses and then decide whether or not to market the batch by balancing those probabilities against the utilities of the possible consequences of the actions, in the manner described by a branch of learning known as Decision Theory. Indeed, this is surely the more plausible account.

[8] Even some vigorous opponents of the Neyman-Pearson method, such as, A.W.F. Edwards (1972, p. 176) accept this defence.

Significance Levels and Inductive Support

The fact that theories are not generally assessed in the black-and-white terms of acceptance and rejection is acknowledged by many classical statisticians, as we see from attempts that have been made to find in significance tests some graduated measure of evidential support. For example, Cramér (1946, p. 421–23) wrote of results being "almost significant", "significant" and "highly significant", depending on the value of the test-statistic. Although he cautiously added that such terminology is "purely conventional", it is clear that he intended to suggest an inverse relationship between the strength of evidence against a null hypothesis and the significance level that would just lead to its rejection.[9]

Indeed, he implies that when this significance level exceeds some (unspecified) value, the evidence ceases to have a negative impact on the null hypothesis and starts to support it; thus, when the χ^2 value arising from some of Mendel's experiments on pea plants was a good way from rejecting Mendel's theory at the 5 percent level, Cramér concluded that "the agreement must be regarded as good", and, in another example, when the hypothesis would only be rejected if the significance level were as high as 0.9, Cramér said that "the agreement is very good".

Classical statisticians commonly try to superimpose this sort of notion of strength of evidence or inductive support on their analyses. For instance, Weinberg and Goldberg (1990, p. 291): "The test result was significant, indicating that H_1 . . . was *a more plausible* statement about the true value of the population mean . . . than [the null hypothesis] H_0". The words we have italicised would, of course, be expected in a Bayesian analysis, but they have no legitimacy or meaning within classical philosophy. And the gloss which the authors then add is not any clearer or better founded: "all we have shown is that *there is reason to believe* that [H_1 is true]". And, of the same result, which was very improbable according to the null hypothesis and significant at the 0.0070

[9] The significance level that, for a given result, would just lead to the rejection of a null hypothesis is also called the p-value of that result, as we stated earlier. "The lower the p-value, the less plausible this null hypothesis . . . and the more plausible are the alternatives" (Wood 2003, p. 134).

level, they say that it is "*quite inconsistent* with the null hypothesis" (*ibid.*, p. 282).

But the result is not "inconsistent" with the null hypothesis, in the logical sense of the term. And as no useful alternative sense seems to exist—certainly none has been suggested—the term in this context is quite misleading. And the project of linking significance levels with strength of evidence has no prospect of success. To prove such a link, you would need to start with an appropriate concept of evidential or inductive support; in fact, no such concept has been formulated in significance test terms, nor is one likely to be. This, for two compelling reasons. First, the conclusions of significance tests often flatly contradict those that an impartial scientist or ordinary observer would draw. Secondly, significance tests depend on factors that it is reasonable to regard as extraneous to judgments of evidential support. We deal with these objections in the next three subsections.

A Well-Supported Hypothesis Rejected in a Significance Test

The first objection was developed in considerable generality by Lindley, 1957, and is sometimes referred to as Lindley's Paradox. We illustrate it with our tulip example. Table 5.6 lists the numbers of red tulips in random samples of size n that would just be sufficient to reject the null hypothesis at the 0.05 level.

It will be noticed that as n increases, the critical proportion of red tulips in the sample that would reject h_1 at the 0.05 level approaches more closely to 40 percent, that is, to the proportion hypothesized in h_1. Bearing in mind that the only alternative to h_1 that the example allows is that the consignment contains red tulips in the proportion of 60 percent, an unprejudiced consideration would clearly lead to the conclusion that as n increases, the supposedly critical values *support* h_1 more and more.

The table also includes information about the power of each test, and shows that the classical thesis that a null hypothesis may be rejected with greater confidence, the greater the power of the test is not borne out; indeed, the reverse trend is signalled.

Freeman (1993, pp. 1446–48) is one of the few to have proposed a way out of these difficulties, without abandoning the

TABLE 5.6

The sample size, n	*The number of red tulips (expressed as a proportion of* n*) that would just reject* h_1 *at the 5% level.*	*The power of the test against* h_2
10	0.70	0.37
20	0.60	0.50
50	0.50	0.93
100	0.480	0.99
1,000	0.426	1.0
10,000	0.4080	1.0
100,000	0.4026	1.0

basic idea of the significance test. He argued that Neyman and Pearson should not have formulated their tests as they did, by first fixing a significance level and then selecting the rejection region that maximizes power. It is this that renders them vulnerable to the Lindley Paradox, because it means that the inductive import of a rejection at a given significance level is the same whatever the size of the sample. Instead, Freeman proposes that the primary role should go to the *likelihood ratio*—that is, the ratio of the probabilities of the data relative to the null and an alternative hypothesis. And he argued that in a significance test, the rule should be to reject the null hypothesis if the likelihood ratio is less than some fixed value, on the grounds that this ensures that the probabilities of *both* the type I *and* the type II errors diminish as the sample size increases.

Freeman's rule is a version of the so-called *Likelihood Principle*, according to which the inductive force of evidence is contained entirely in the likelihood ratios of the hypotheses under consideration. This principle, in fact, follows directly from Bayes's theorem (see Section 4.c) and is unavoidable in Bayesian inductive inference. Freeman (1993, p. 1444) too regards this principle as essential—"the one secure foundation for all of statistics"—but

neither he nor any other non-Bayesian has proved it. And this is not surprising, for they strenuously deny that hypotheses have probabilities, and it is precisely upon this idea that the Bayesian proof depends. The likelihood principle therefore cannot save significance tests from the impact of Lindley's Paradox, which, it seems to us, shows unanswerably and decisively that inferences drawn from significance tests have no inductive significance whatever.

We now consider a couple more aspects of significance tests which reinforce this same point.

The Choice of Null Hypothesis

In a Neyman-Pearson test you need to choose which of the competing hypotheses to treat as the null hypothesis, and the result of that choice has a bearing on which is finally accepted and which rejected. Take the tulip example again: if an experiment showed 50 red-flowering plants in a random sample of 100, then h_1 (40 percent red) would be rejected at the 0.05 level if it were the null hypothesis, and h_2 (60 percent red) would be accepted. But with h_2 as null hypothesis, the opposite judgment would be delivered! It will be recalled that the role of null hypothesis was filled by considering the desirability, according to a personal scale of values, of certain practical consequences of rejecting a true hypothesis; and where the hypotheses were indistinguishable by this practical yardstick, the null hypothesis could be designated arbitrarily. But pragmatic and arbitrary decisions such as these have no epistemic meaning and cannot form the basis of inductive support.

Another sort of influence on significance tests that is also at odds with their putative role in inductive reasoning arises through the stopping rule.

The Stopping Rule

Significance tests are performed by comparing the probability of the outcome obtained with the probabilities of other possible outcomes, in the ways we have described. Now the space of possible outcomes is created, in part, by what is called the *stopping rule*; this is the rule that fixes in advance the circumstances under

which the experiment should stop. Our trial to test the fair-coin hypothesis, for example, was designed to stop after the coin had been flipped 20 times. Another stopping rule for that experiment might have instructed the experimenter to end it as soon as 6 heads appeared, which would exclude many of the outcomes that were previously possible and introduce an infinity of new ones. Expressed as the number of heads and tails in the outcome, the possibilities for the two stopping rules are: (*20,0*), (*19,1*), . . . , (*0,20*), in the first case, and (*6,0*), (*6,1*), (*6,2*), . . . , and so on, in the second. The two stopping rules have surprisingly and profoundly different effects.

Consider, for example, the result (6,14), which could have arisen with either stopping rule. When the rule was to stop after 6 heads, the null hypothesis would be rejected at the 0.05 level. This is shown as follows: the assumed stopping rule produces the result (*6, i*) whenever (*5, i*), appearing in any order, is then succeeded by a head. Thus, relative to the fair-coin hypothesis, the probability of the result (*6, i*) is given by ${}^{i+5}C_5(\frac{1}{2})^5(\frac{1}{2})^i \times \frac{1}{2}$. Table 5.7 shows the sampling distribution.

TABLE 5.7

The Probabilities of Obtaining *i* Tails with a Fair Coin in a Trial of Designed to Stop after 6 Heads Appear.

Outcome (H,T)	*Probability*	*Outcome (H,T)*	*Probability*
6,0	0.0156	6,11	0.0333
6,1	0.0469	6,12	0.0236
6,2	0.0820	6,13	0.0163
6,3	0.1094	6,14	0.0111
6,4	0.1230	6,15	0.0074
6,5	0.1230	6,16	0.0048
6,6	0.1128	6,17	0.0031
6,7	0.0967	6,18	0.0020
6,8	0.0786	6,19	0.0013
6,9	0.0611	16,20	0.0008
6,10	0.0458	6,21	0.0005
		etc.	etc.

We see from the table that the results which are at least as improbable as the actual one are (*6,14*), (*6,15*), . . . , and so on, whose combined probability is 0.0319. Since this is below the critical value of 0.05, the result (*6,14*) is significant at this level and the null hypothesis should therefore be rejected. It will be recalled that when the stopping rule predetermined a sample size of 20, the very same result was not significant.[10] So in calculating the significance of the outcome of any trial, it is necessary to know the stopping rule that informed it.

We have considered just two stopping rules that could have produced some particular result, but any number of others have that same property. And not all of these other possibilities rest the decision to stop on the outcomes themselves, which some statisticians regard as not quite legitimate. For instance, suppose that after each toss of the coin, you drew a playing card at random from an ordinary pack, with the idea of calling the trial off as soon as the Queen of Spades has been drawn. This stopping rule introduces a new outcome space, which will lead to different conclusions in certain cases. Or suppose the experimenter intends to continue the trial until lunch is ready: in this case, the sampling distribution could only be worked out with complex additional information about the chance, at each stage of the trial, that preparations for the meal are complete.

The following example brings out clearly how inappropriate it is to involve the stopping rule in the inductive process: two scientists collaborate in a trial, but are privately intent on different stopping rules; by chance, no conflict arises, as the result satisfies both. What then are the outcome space and the sampling distribution for the trial? To know these you would need to discover how each of the scientists would have reacted in the event of a disagreement. Would they have conceded or insisted, and if they had put up a fight, which of them would have prevailed? *We suggest that such information about experimenters' subjective intentions, their physical strengths and their personal qualities has no inductive relevance whatever in this context, and that in practice it is never sought or even contemplated. The fact that significance tests and, indeed, all classical inference models require it is a decisive objection to the whole approach.*

[10] This illustration of the stopping-rule effect is adapted from Lindley and Phillips 1976.

Whitehead (1993) is one of the few to have defended the stopping rule as an essential component of the inductive process. He denies that the subjective intention underlying the stopping rule is irrelevant, illustrating his point with a football match, of all things, in which the captain of one side is allowed to decide when the game should finish, and in fact blows the whistle when his team is 1–0 ahead. Whitehead remarks that learning the stopping rule here would reduce his high opinion of the winning side. To revert to a case where classical statistics can more obviously be applied, this is analogous to an experimenter, who is predisposed in favour of one of the hypotheses, deciding to stop sampling as soon as more red than yellow tulips have flowered. If the final count were, say, 1 red and 0 yellow, we would indeed not be much swayed in favour of the experimenter's preferred hypothesis, but not because of the known bias, or the stopping rule, rather, we suggest, because of the smallness of the sample. To believe otherwise runs into the objection we raised earlier, namely, that if the biased experimenter were working with an impartial, or differently biased colleague, who was actuated by a different stopping rule, you would have to delve into the personal qualities of the experimenters in order to discover the outcome space of the experiment, and hence the inductive significance of the result.

Experimenters' prejudices can only have inductive significance for us if we believe them to have clairvoyant knowledge about future samples; but this is just what a random sampling experiment effectively precludes. On the other hand, the captain in charge of the stopping rule in the hypothetical football match does have information about the likely course of the game, since he may know the teams' recent form and can observe how well each side is presently playing. But a football game is not a random sampling experiment, and is therefore an unsuitable example in this context.

Gillies (1990, p. 94) also argued that the stopping rule is an essential part of a scientific inference. He claimed that "to those who adopt falsificationism (or a testing methodology)" it "seems natural and only to be expected" that the stopping rule should in general affect a theory's empirical support, because "wherever

possible the experimental method should be applied, and this consists in designing and carrying out a *repeatable* experiment . . . whose result might refute h [the null hypothesis]". This, he claims, means that the stopping rule is evidentially relevant.

In response, we certainly concede that it can do no harm and might do good to repeat an experiment. But why should it be repeat*able*? Many useful and informative tests cannot be repeated: for example, pre-election opinion polls and certain astronomical observations. Would our confidence in the age of the Turin Shroud be any different if the entire cloth had been consumed in the testing process, so precluding further tests? Surely not.

Moreover, in an important sense, no experiment is repeatable, for none could ever be done in exactly the same way again. Indefinitely many factors alter between one performance of an experiment and another. Of course, not all such changes matter. For instance, the person who tossed the coin might have worn yellow shoes or sported a middle parting; but these are irrelevant, and if you called for the experiment to be repeated, you would issue no instructions as to footwear or hairstyle. On the other hand, whether or not the coin had a piece of chewing gum attached to one side, or a strong breeze was blowing when it was tossed should be taken into account. The question then is whether the stopping rule falls into the first category of irrelevant factors or into the second of relevant ones. Gillies (*ibid.*, p. 94) simply presumes the latter, arguing, with reference to the coin trial, that the "test of h in this case consists of the whole carefully designed experimental procedure", and suggesting thereby that this procedure must include reference to the stopping rule. But Gillies neither states this explicitly nor provides any reason why it should be so—unavoidably, in our view.

We show in Chapter 8 that in the Bayesian scheme the posterior probabilities in each case are unaffected by the subjective intentions implicit in the stopping rules and depend on the result alone. Thus, if the experimental result is, for instance, *6 heads, 14 tails*, it does not matter whether the experimenter had intended to stop the trial after 20 tosses of the coin, or after *6 heads*, or after lunch, or after the Queen of Spades has made her entrance, or whatever.

5.e | Testing Composite Hypotheses

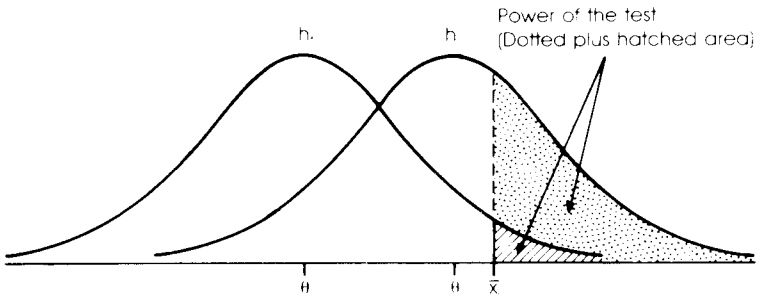
We have, so far, restricted our account of the Neyman-Pearson method to cases where just two, specific hypotheses are assumed to exhaust the possibilities. But such cases are atypical in practice, and we need to look at how Neyman and Pearson extended and modified their approach to deal with a wider range of alternative hypotheses. They considered, for instance, how to test a hypothesis, h_1, that some population parameter, θ, has a specific value, say θ_1, against the unspecific, composite hypothesis, h_2, that $\theta > \theta_1$.

The principle of maximizing the power of a test for a given significance level cannot be applied where these are the competing hypotheses. For, although the situation allows one to determine a critical region corresponding to any designated significance level, as before, the probability of a type II error is indeterminate. Neyman and Pearson responded by varying their central principle.

They first of all proposed that in such cases one should choose for the critical region one that has maximum power for each component of h_2. Tests satisfying this new criterion are called *Uniformly Most Powerful* (UMP). But they ran into the problem that, relative to some elements of h_2, the power of UMP tests might be very low, indeed, lower than the significance level, in which case there would be a greater chance of rejecting the null hypothesis when it is true than when it is false. This possibility was unacceptable to Neyman and Pearson and to avoid it, they imposed the further restriction that the test should be *unbiased*, that is, its power relative to each element of h_2 should be at least as great as its significance level.

We can illustrate the idea of tests that are both UMP and unbiased (UMPU) with a simple example. The test will be based on the mean, $\bar{x}$, of a random sample drawn from a normal population with known standard deviation and unknown mean, θ. The diagram shows the sampling distributions relative to the null hypothesis h_1: $\theta = \theta_1$, and relative to an arbitrary element, h_i, of the composite alternative hypothesis h_2: $\theta > \theta_1$.

AN UMPU TEST



Consider a critical region for rejecting h_1 consisting of points to the right of the critical value, $\bar{x}_c$. The area of the hatched portion is proportional to the significance level. The area under the h_i-curve to the right of $\bar{x}_c$ represents the probability of rejecting h_1 if h_i is true (and hence, h_1 false). Clearly, the closer are the means specified by h_1 and h_i, the smaller this probability will be. But it could never be less than the significance level. In other words, the test is UMPU.

Suppose now that the range of alternatives to the null hypothesis is greater still and that h_1: $\theta = \theta_1$ is to be tested against h_2: $\theta \neq \theta_1$, the standard deviation again being known. A critical region located in one tail of the sampling distribution associated with h_1 would not now constitute an unbiased test. But an UMPU test can be constructed by dividing the critical region equally between the two tails. This can be appreciated diagrammatically and also rigorously shown (see for instance Lehmann 1986). An UMPU test thus provides some basis for the two-tailed tests implied by Fisher's theory, for which he offered no rationale.

But UMPU tests are in fact rather academic, since they exist in very few situations. And they depart somewhat from the Neyman-Pearson ideal of maximum power for a given significance level, in that the power of such a test can never be determined; so in particular cases, it may, for all we know, be only infinitesimally different from the significance level. More seriously, the modifications introduced to meet the challenge of composite hypotheses are equally afflicted by the various difficulties we have shown to discredit even the most uncomplicated form of the significance test.

5.f Classical Estimation Theory

Scientists often estimate a physical quantity and come thereby to regard a certain number, or range of numbers, as a more or less good approximation to the true value. Significance tests do not in general deliver such estimates and the need for them has prompted classical statisticians to develop a distinct body of doctrine known as Estimation Theory. The theory is classical, in that it purports to provide objective, non-probabilistic conclusions. It has two aspects, namely, point estimation and interval estimation, both of which we regard as fallacious, as we explain in the following review.

5.f.1. Point Estimation

Point estimation differs from interval estimation, which we deal with below, in offering a single number as the so-called 'best estimate' of a parameter. Suppose a population parameter, such as its mean, is in question. The technique for estimating this is to draw a random sample of predetermined size, n, from the population, and to measure each element drawn. Then, letting $x = x_1, \ldots, x_n$ denote the measurements thus derived, the next step is to pick an estimating statistic, t, this taking the form of a calculable function $t = f(x)$. Finally, the best estimate of the unknown parameter is inferred to be t_0, the value of t yielded by the experiment. But not every statistic is accepted as an estimator. The authors of this approach to estimation have specified certain conditions that any estimator must meet; the most frequently mentioned being the conditions of *sufficiency*, *unbiasedness*, *consistency* and *efficiency*. We discuss these in turn.

Sufficient Estimators

It will be recalled from Section 5.b, that a statistic t is sufficient for θ when $P(x \mid t)$ is independent of θ. The present requirement is that any estimating statistic should be sufficient in this sense. The mean of a random sample satisfies the requirement when it is

used to estimate a population mean, but the sample range, for example, is not. Nor is the sample median.[11]

Sufficiency is a Bayesian requirement too. Expressed in Bayesian terms a statistic, t, is sufficient for θ, just in case $P(x \mid t) = P(x \mid t \ \& \ \theta)$, for all θ. It follows straightforwardly from Bayes's theorem that t is sufficient for θ if, and only if, $P(\theta \mid t) = P(\theta \mid x)$. Hence, when a sample statistic is sufficient, it makes no difference whether you calculate the posterior distribution using it or using the full experimental information in x; the results will be the same. In other words, a sufficient statistic contains all the information that in Bayesian terms is relevant to θ.

A compelling intuition tells us that, in evaluating a parameter, we should not neglect any relevant information. There is a satisfactory rationale for this. Suppose you have two bits of information, a and b. There are then three posterior distributions to consider: $P(\theta \mid a)$, $P(\theta \mid b)$ and $P(\theta \mid a \ \& \ b)$. If these differ, which should describe your current belief state? The Bayesian has no choice, for $P(\theta \mid a)$ is your distribution of beliefs were you to learn a and nothing else. But you have in fact learned $a \ \& \ b$ and nothing else. Therefore, your current belief state must be described by $P(\theta \mid a \ \& \ b)$, rather than by the other mathematical possibilities.

The injunction to use all the relevant evidence in an inductive inference, which Carnap (1947) called the *Total Evidence Requirement*, is often considered to be an independent postulate. This is true, at any rate, within classical estimation theory, which also has to rely on the intuition, which it cannot prove either, that a sufficient statistic captures all the relevant evidence. The intuitions are well founded, but their source, in our opinion, is Bayes's theorem, applied unconsciously.

Unbiased Estimators

These are defined in terms of the expectation, or expected value, of a random variable, which is given by $E(x) = \Sigma x_i P(x_i)$, the sum

[11] The sample *range* is the difference between the highest and lowest measurements; if the sample measurements are arranged in increasing order, and if n is odd, their *median* is the middle element of the series; if n is even, the median is the higher of the middle two elements.

or in the continuous case, the integral, being taken over all possible values of x_i. We mentioned in Chapter 2 that the expectation of a random variable is also called the *mean* of its probability or density distribution; when the distribution is symmetrical, its mean is also its geometric centre. A statistic is defined as an *unbiased estimator* of θ just in case its expectation equals the parameter's true value. The idea is often glossed by saying that the value of an unbiased statistic, averaged over repeated samplings, will "in the long run", be equal to the parameter being estimated.[12]

Many intuitively satisfactory estimators are unbiased, for instance the proportion of red counters in a random sample is unbiased for the corresponding proportion in the urn from which it was drawn, and the mean of a random sample is an unbiased estimator of the population mean. However, sample variance is not an unbiased estimator of population variance and is generally "corrected" by the factor $\frac{n}{n-1}$.

But unbiasedness is neither a necessary nor a sufficient condition for a satisfactory estimation. We may see this through an example. Suppose you draw a sample, of predetermined size, from a population and note the proportion of individuals in the sample with a certain trait, and at the same time, you toss a standard coin. We now posit an estimating statistic which is calculated as the sample proportion plus k (> 0), if the coin lands heads, and plus k', if it lands tails. Then, if $k = -k'$, the resulting estimator is unbiased no less than the sample proportion itself, but its estimates are very different and are clearly no good. If, on the other hand, $k' = 0$, the estimator is biased, yet, on the occasions when the coin lands tails, the estimates it gives seem perfectly fine.

Not surprisingly, then, one finds the criterion defended, if at all, in terms which have nothing to do with epistemology. The usual defence is concerned rather with pragmatics. For example, Barnett claimed that, "within the classical approach unbiasedness is often introduced as a *practical* requirement to limit the class of estimators" (1973, p. 120; our italics). Even Kendall and Stuart, who wrote so confidently of the need to correct biased

[12] See for example Hays 1963, p. 196.

estimators, conceded that they had no epistemic basis for this censorious attitude:

> There is *nothing except convenience* to exalt the arithmetic mean above other measures of location as a criterion of bias. We might *equally well* have chosen the median of the distribution of t or its mode as determining the "unbiased" estimator. The mean value is used, as always, *for its mathematical convenience*. (1979, p. 4; our italics)

These authors went on to warn their readers that "the term 'unbiased' should not be allowed to convey overtones of a non-technical nature". But the tendentious nature of the terminology makes such misleading overtones hard to avoid. The next criterion is also named in a way that promises more than can be delivered.

Consistent Estimators

An estimator is defined to be *consistent* when, as the sample size increases, its probability distribution shows a diminishing scatter about the parameter's true value. More precisely, a statistic derived from a random sample of size n is a consistent estimator for θ if, for any positive number, ε, $P(|t - \theta| \leq \varepsilon)$ tends to 1, as n tends to infinity. This is sometimes described as t tending probabilistically to θ.

There is a problem with the consistency criterion as described, because it admits estimators that are clearly inadmissible. For example, if T_n is a consistent estimator, so is the estimator, T_n', defined as equal to zero for $n \leq 10^{10}$ and equal to T_n for $n > 10^{10}$. Fisher therefore added the further restriction that an admissible estimator should, in Rao's words, be "an explicit function of the observed proportions only". So, if the task is to estimate a population proportion, θ, the estimator should be a consistent function just of the corresponding sample proportion, and it should be such that when the observed and the population proportions happen to coincide, the estimator gives a true estimate (Rao 1965, p. 283). This adjustment appears to eliminate the anomalous estimators.

Fisher believed that consistency was the "fundamental criterion of estimation" (1956, p. 141) and that non-consistent estimators "should be regarded as outside the pale of decent usage" (1970, p. 11). In this, Neyman (1952, p 188) agreed "perfectly" with Fisher, and added his opinion that "it is definitely not profitable to use an inconsistent estimate."[13] Fisher defended his emphatic view in the following way:

> as the samples are made larger without limit, the statistic will usually tend to some fixed value characteristic of the population, and, therefore, expressible in terms of the parameters of the population. If, therefore, such a statistic is to be used to estimate these parameters, there is only one parametric function to which it can properly be equated. If it be equated to some other parametric function, we shall be using a statistic which even from an infinite sample does not give a correct value. . . . (1970, p. 11)

Fisher's claim here is that because a consistent estimator converges to some parameter value, it "can properly be equated" to that value, and it should be equated to no other value, because, if the sample were infinite, it would then certainly give the wrong result. This is more assertion than argument and in fact is rather implausible in its claims. Firstly, one should not, without qualification, equate an unknown parameter with the value taken by a statistic in a particular experiment; for, as is agreed on all sides, such estimates may, almost certainly will be in error, a consideration that motivates interval estimation, which we discuss below. Secondly, the idea that a consistent estimator becomes more accurate as the sample increases, and perfectly so in the limit, implies nothing at all about its accuracy on any particular occasion. Arguing for an estimator with this idea in mind would be like defending the use of a dirty measuring instrument on the grounds that if it were cleaner it would be better; in assessing a result, we need to know how good the instrument was in the experiment at hand, not how it might have performed under different conditions.

And (rebutting Fisher's last point) just as estimates made by consistent estimators may be quite inaccurate and clearly wrong, those from non-consistent ones might be very accurate and

[13] Presumably this should read: 'estimator'.

clearly right. For instance, suppose $\bar{x} + (n - 100)\bar{x}$ were chosen to estimate a population mean. This odd statistic is non-consistent, for, as the sample size grows, it diverges ever more sharply from the population mean. Yet for the special case where $n = 100$, the statistic is equivalent to the familiar sample mean, and gives an intuitively satisfactory estimate.

Efficient Estimators

The above criteria are clearly incomplete, because they do not incorporate the obvious desideratum that an estimate should improve as the sample becomes larger. So, for instance, a sample mean that is based on a sample of 2 would be 'sufficient', 'unbiased' and 'consistent', yet estimates of the population mean derived from it would not inspire confidence, certainly not as much as when there are 100, say, in the sample. This consideration is addressed by classical statistics through the efficiency criterion: the smaller an estimator's variance about the parameter value, the more *efficient* it is said to be, and the better it is regarded. And since the variance of a sample statistic is generally inversely dependent on the size of the sample, the efficiency criterion reflects the preference for estimates made with larger samples.

But it is not easy to establish, in classical terms, why efficiency should be a measure of quality in an estimator. Fisher (1970, p. 12) stated confidently that the less efficient of two statistics is "definitely inferior . . . in its accuracy"; but since he would have strayed from classical principles had he asserted that particular estimates were certainly or probably correct, even within a margin of error, this claim has no straightforward meaning. Kendall and Stuart's interpretation is the one that is widely approved. A more efficient statistic, they argued, will "deviate less, on the average, from the true value" and therefore, "we may reasonably regard it as better" (1979, p. 7). Now it is true that if e^i_1 and e^i_2 are the estimates delivered by separate estimators on the ith trial and if θ is the true value of the parameter, then there is a calculable probability that $| e^i_1 - \theta | < | e^i_2 - \theta |$, which will be greater the more efficient the first estimator is than the second. Kendall and Stuart translate this probability into an average frequency in a long run of trials, which, as we already remarked, goes beyond

logic. But even if the translation were correct, the performance of an estimator over a hypothetical long run implies nothing about the closeness of a particular estimate to the true value. And since estimates are usually expensive and troublesome to obtain and often inform practical actions, what is wanted and needed are just such evaluations of particular estimates.

5.f.2. Interval Estimation

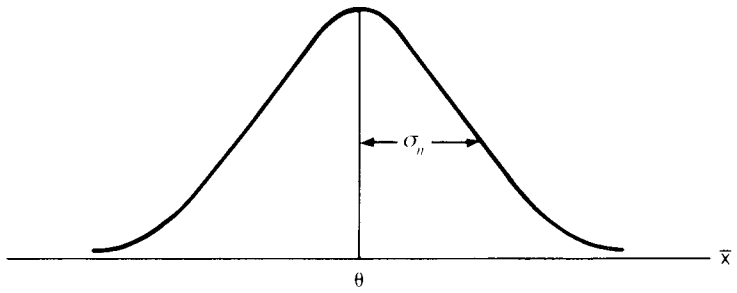
In practice, this demand is evidently met, for estimates are normally presented as a range of numbers, for example, in the form $\theta = a \pm b$, not as point values, which, as Neyman observed, "it is more or less hopeless to expect . . . will ever be equal to the true value" (1952, p. 159). Bayesians would qualify an interval estimate by the subjective probability that it contains the true value (see the discussion of 'credible intervals' in Section 8.a). Neyman's theory of confidence intervals, developed around 1930, and now dominant in the field, was intended to give a classical expression to this idea.

Confidence Intervals

Consider the task of estimating the mean height, θ, of the people in a large population, whose standard deviation, σ, is known. A sample of some predetermined size, n, is randomly selected, and its mean, $\bar{x}$, is noted. This mean can take many possible values, some more probable than others; the distribution representing this situation is approximately normal (the larger the population, the closer the approximation) with a mean equal to that of the population and a standard deviation given by $\sigma_n = \sigma n^{-\frac{1}{2}}$.

The sampling distribution plots possible sample means against probability densities, not probabilities, and, as explained earlier, this signifies that the probability that $\bar{x}$ lies between any two points is proportional to the area enclosed by those points and the curve. Because the distribution is essentially normal, it follows that, with probability 0.95,

$$-1.96\,\sigma_n \leq \theta - \bar{x} \leq 1.96\,\sigma_n.$$



The sampling distribution of means from a population with mean θ and standard deviation σ.

And this implies that, with probability 0.95,

$$\bar{x} - 1.96\,\sigma_n \leq \theta \leq \bar{x} + 1.96\,\sigma_n.$$

Let m be the value of $\bar{x}$ in a particular experimental sample; since θ and n are known, the terms $m - 1.96\,\sigma_n$ and $m + 1.96\,\sigma_n$ can be computed. The interval between these two values is called a *95 percent confidence interval* for θ. Clearly there are other confidence intervals relating to different regions of the sampling distribution, and others, too, associated with different probabilities. The probability associated with a particular confidence interval is called its *confidence coefficient*.

The probability statements given above are simply deductions from the assumptions made and are unquestionably correct; and what we have said about confidence intervals, being no more than a definition of that concept, is also uncontroversial. Controversy arises only when confidence intervals are assigned inductive meaning and interpreted as estimates of the unknown parameter. What we shall call the *categorical-assertion* interpretation and the *subjective-confidence* interpretation are the two main proposals for legitimizing such estimates. We deal with these in turn.

The Categorical-Assertion Interpretation

This interpretation was first proposed by Neyman and has been widely adopted. Neyman said (1937, p. 263) that the "practical statistician", when estimating a parameter, should calculate a confidence interval and then "state" that the true value lies between the two confidence bounds, in the knowledge that (when the confidence coefficient is 0.99) "in the long run he will be correct in about 99 percent of all cases". The statistician's statement should not signify a belief in its truth, however. Indeed, Neyman (1941, p. 379) rejected the very idea of reasoning inductively to a conclusion, because he believed that "the mental process leading to knowledge . . . can only be deductive". Induction, for Neyman, was rather a matter of behaviour, and in the case of interval estimates, the proper outcome was a decision "to behave as if we actually knew" that the parameter lies within the confidence bounds.

We have already discussed this interpretation (in Section 5.d) in the context of significance tests and argued that typically and more reasonably scientists evaluate theories by degree; they do not, and, moreover, should not act in the way that Neyman recommended. A further indication that the interpretation is wrong arises from the fact that confidence intervals are not unique, as we explain next.

Competing Intervals

It is obvious from the sampling distribution of means depicted above that indefinitely many regions of that normal distribution cover 95 percent of its area. So instead of the usual 95 percent confidence interval located at the centre of the distribution, one could consider asymmetrical confidence intervals or ones that extend further into the tails while omitting smaller or larger strips in the centre. Neyman's categorical-assertion interpretation requires one to "assert" and "behave as if one actually knew" that the parameter lies in each and every one of this multiplicity of possible 95 percent confidence intervals, which is clearly unsatisfactory, and indeed, paradoxical.

Defenders of the interpretation have reacted in two ways, both we believe unsatisfactory. The first discriminates between confidence intervals on the basis of their length, and claims that, for a given confidence coefficient, the shortest interval provides the best estimate. In the words of Hays (1969, p. 290), there is "naturally

. . . an advantage in pinning the population parameter within the narrowest possible range with a given probability". By this criterion, the centrally symmetrical interval $m \pm 1.96\sigma_{\bar{x}}$ is the preferred 95 percent confidence interval for the population mean in the example cited above. The preference is based on the idea that the width of a confidence interval is a measure of the 'precision' of the corresponding estimate, and that this is a desirable feature. Thus Mood (1950, p. 222), when comparing two 95 percent confidence intervals, stated that the longer one was inferior, "for it gives less precise information about the location" of the parameter.

But it is not true that the length of a confidence interval measures its precision. For, consider the interval $| a, b |$ as an estimate of θ, and the interval $| f(a), f(b) |$ as an estimate of $f(\theta)$. If f is a 1-1 function, the two estimates are equivalent and must be equally informative and therefore equally precise. But while the first may be the shortest 95 percent confidence interval for θ, the second might not be the shortest such interval for $f(\theta)$; this would be the case, for instance, when $f(a) = a^{-1}$.

Another difficulty is that different sample statistics may yield different minimum-length confidence intervals, a fact that has prompted the proposal to restrict interval estimates to those given by statistics with the smallest possible variance. It is argued that although this new criterion does not guarantee the shortest possible confidence interval in any particular case, it does at least ensure that such intervals "are the shortest on average in large samples" (Kendall and Stuart 1979, p. 126). We have already criticized both the long-run justification and the short-length criterion, and since two wrongs don't make a right, we shall leave the discussion there.

Neyman (1937, p. 282) suggested another way of discriminating between possible confidence intervals. He argued, in a manner familiar from his theory of testing, that a confidence interval should not only have a high probability of containing the correct value but should also be relatively unlikely to include wrong values. More precisely: a best confidence interval, I_o, should be such that for any other interval, I, corresponding to the same confidence coefficient, $P(\theta' \in I_o \mid \theta) \leq P(\theta' \in I \mid \theta)$; moreover, the inequality must hold whatever the true value of the parameter, and for every value θ' different from θ. But as Neyman himself showed, there are no 'best' intervals of this kind for most of the cases with which he was originally concerned.

The Subjective-Confidence Interpretation

Neyman's categorical-assertion interpretation, contrary to its main intention, does, in fact, contain an element that seems to imply a scale by which estimates may be evaluated and qualified, namely the confidence coefficient. For suppose some experimentally established range of numbers constituted a 90 percent confidence interval for θ, rather than the conventionally approved 95 or 99 percent, we would still be enjoined to assert that θ is in that range and to act as if we believed that to be true, though with a correspondingly modified justification that now referred to a 90 percent frequency of being correct "in the long run". But if the justification had any force (we have seen that it does not), it would surely be stronger the lower the frequency of error. So the categorical assertion that θ is in some interval must after all be qualified by an index running from 0 to 100 indicating how well founded it is, and this is hard to distinguish from the index of confidence that is explicit in the subjective-confidence interpretation that we deal with now.

In this widely approved position, a confidence coefficient is taken to be "a measure of our confidence" in the truth of the statement that the confidence interval contains the true value (for example, Mood 1950, p. 222). This has some surface plausibility. For consider again the task of estimating a population mean, θ. We know that in experiments of the type described, θ is included in any 95 percent confidence interval with an objective probability of 0.95; and this implies that if the experiment were performed repeatedly, θ would be included in such intervals with a relative frequency that tends, in the limit, to 0.95. It is tempting, and many professional statisticians find it irresistible, to infer from this limit property a probability of 0.95 that the particular interval obtained in a particular experiment does enclose θ. But drawing such an inference would commit a logical fallacy.[14]

[14] This fallacy is repeated in many statistics texts. For example, Chiang 2003, (pp. 138–39): "The probability that the interval contains μ is either zero or one;

The subjective-confidence interpretation seems to rely on a misapplication of a rule of inference known as the Principle of Direct Probability (see Chapter 3), which is used extensively in Bayesian statistics. The principle states that if the objective, physical probability of a random event (in the sense of its limiting relative frequency in an infinite sequence of trials) is known to be r, then, in the absence of any other relevant information, the appropriate subjective degree of belief that the event will occur on any particular trial is also r. Expressed formally, if the event in question is a, and $P^*(a)$ is its objective probability, and if a_t describes the occurrence of the event on a particular trial, the Principle of Direct Probability says that $P[a_t \mid P^*(a) = r] = r$, where P is a subjective probability function.

For example, the physical probability of getting a number of heads, K, greater than 5 in 20 throws of a fair coin is 0.86 (see Table 5.1 above), that is, $P^*(K > 5) = 0.86$. By the Principle of Direct Probability,

$$P[(K > 5)_t \mid P^*(K > 5) = 0.86] = 0.86.$$

That is to say, 0.86 is the confidence you should have that any particular trial of 20 throws of a fair coin will produce more than 5 heads. Suppose one such trial produced 2 heads. To infer that we should now be 86 percent confident that 2 is greater than 5 would, of course, be absurd; it would also be a misapplication of the principle. For one thing, if it were legitimate to substitute numbers for K, why would such substitution be restricted to its first occurrence in the principle? But in fact, no such substitution is allowed. For the above equation does not assert a general rule for each number K from 0 to 20; the K-term is not a number, but a function that

no intermediate values are possible. What then is the initial probability of 0.95? Suppose we take a large number of samples, each of size n. For each sample we make a statement that the interval observed from the sample contains μ. Some of our statements will be true, others will not be. According to [the equation we give in the text, above] . . . 95 percent of our statements will be true. In reality we take only one sample and make only one statement that the interval contains μ. Thus [*sic*] we do have confidence in our statement. The measure of our confidence is the initial probability 0.95."

takes different values depending on the outcome of the underlying experiment.

Mistaking this appears to be the fallacy implicit in the subjective-confidence interpretation. It is true that the objective probability of θ being enclosed by experimentally determined 95 percent confidence intervals is 0.95. If I_1 and I_2 are variables representing the boundaries of such confidence intervals, the Principle of Direct Probability implies that

$$P[(I_1 \leq \theta \leq I_2)_t \mid P^*(I_1 \leq \theta \leq I_2) = 0.95] = 0.95,$$

and this tells us that we should be 95 percent confident that any sampling experiment will produce an interval containing θ. Suppose now that an experiment that was actually performed yielded I'_1 and I'_2 as the confidence bounds; the subjective-confidence interpretation would tell us to be 95 percent confident that $I'_1 \leq \theta \leq I'_2$. But this would commit exactly the same fallacy as we exposed in the above counter-example. For I_1 and I_2, like K, are functions of possible experimental outcomes, not numbers, and so the desired substitution is blocked.[15]

In response, it might be said that the subjective-confidence interpretation does not depend on the Principle of Direct Probability (a Bayesian notion, anyway), that it is justified on some other basis. But we know of none, nor do we think any is possible, because, as we shall now argue, the interpretation is fundamentally flawed, since it implies that one's confidence in a proposition should depend on information that is manifestly irrelevant, namely, that concerning the stopping rule, and should be independent of prior information that is manifestly relevant. We address these two points next.

The Stopping Rule

Confidence intervals arise from probability distributions over spaces of possible outcomes. Although one of those outcomes

[15] Howson and Oddie 1979 pointed out this misapplication of the principle in another context. See also 3.f above.

will be actualized, the space as a whole is imaginary, its contents depending in part on the experimenters' intentions, embodied in the adopted stopping rule, as we explained earlier. Estimating statistics employed in point estimation are also stopping-rule dependent, because the stopping rule dictates whether or not those statistics satisfy the various conditions that are imposed on estimators. So, for instance, the sample mean is an unbiased estimator of a population mean if it is based on a fixed, predetermined sample size, but not necessarily otherwise.[16]

The criticism we levelled at this aspect of classical inference in the context of tests of significance applies here too, and we refer the reader back to that discussion. In brief, the objection is that having to know the stopping rule when drawing an inference from data means that information about the experimenters' private intentions and personal capacities, as well as other intuitively extraneous facts, is ascribed an inductive role that is highly inappropriate and counter-intuitive. This is, in a way, tacitly acknowledged by most classical statisticians, who in practice almost always ignore the stopping rule and standardly carry out any classical analysis *as if* the experiment had been designed to produce a sample of the size that it did, without any evidence that this was so, and even when it clearly was not.

We take up the discussion of the stopping rule again in Chapter 8, where we show why it plays no role in Bayesian induction.

Prior Knowledge

Estimates are usually made against a background of partial knowledge, not in a state of complete ignorance. Suppose, for example, you were interested in discovering the average height of students attending the London School of Economics. Without being able to point to results from carefully conducted studies, but on the basis of common sense and what you have learned informally about students and British universities' admission standards, you would feel pretty sure that this could not be below four feet, say, nor above six. Or you might already have made an

[16] See, for example, Lee 1989, p. 213.

exhaustive survey of the students' heights, lost the results and be able to recall with certitude only that the average was over five feet. Now if a random sample, by chance, produced a 95 percent confidence interval of 3' 10" ± 2", you would be required by classical principles to repose an equivalent level of confidence in the proposition that the students' average height really does lie in that interval. But with all you know, this clearly would not be a credible or acceptable conclusion.

A classical response to this difficulty might take one of two forms, neither adequate, we believe. The first would be to restrict classical estimation to cases where no relevant information is present. But this proposal is scarcely practicable, as such cases are rare; moreover, although a little knowledge is certainly a dangerous thing, it would be odd, to say the least, if it condemned its possessor to continue in this condition of ignorance in perpetuity. A second possibility would be to combine in some way informal prior information with the formal estimates based on random samples. The Bayesian method expresses such information through the prior distribution, which then contributes to the overall conclusion in a regulated way, but there is no comparable mechanism within the confines of classical methodology.

5.g | Sampling

Random Sampling

The classical methods of estimation and testing that we have been considering purport to be entirely objective, and it is for this reason that they call for the sampling distribution of the data also to be objective. To this end, classical statisticians require the sample that is used for the estimate to have been generated by an impartial, physical process that ensures for each element of the population an objectively equal chance of being selected. Here is a simple instance of such a process: a bag containing similar counters, each corresponding to a separate member of the population, and marked accordingly, is shaken thoroughly and a counter selected blindfold; this selection is repeated the prescribed number of times, and the population members picked out by the

selected counters then constitute a random sample. There are of course other, more sophisticated physical mechanisms for creating random samples.

What we call the *Principle of Random Sampling* asserts that satisfactory estimates can only be obtained from samples that are objectively random in the sense indicated.

Judgment Sampling

The Principle of Random Sampling may be contrasted with another approach, which is motivated by the wish to obtain a *representative sample*, one that resembles the population in all those respects that are correlated with the characteristic being measured. Suppose the aim were to measure the proportion of the population intending to vote Conservative in a forthcoming election. If, as is generally agreed, voting preference is related to age and socio-economic status, a representative sample should recapitulate the population in its age and social class structure; quite a number of other factors, such as gender and area of residence, would, no doubt, also be taken into account in constructing such a sample. A representative sample successfully put together in this way will have the same proportion of intending Conservative voters as the parent population. Samples selected with a view to representativeness are also known as *purposive*, or *judgment samples*; they are not random. A kind of judgment sampling that is frequently resorted to in market research and opinion polling is known as *quota sampling*, where interviewers are given target numbers of people to interview in various categories, such as particular social classes and geographical regions, and invited to exercise their own good sense in selecting representative groups from each specified category.

Some Objections to Judgment Sampling

Judgment sampling is held to be unsatisfactory by many statisticians, particularly those of a classical stripe, who adhere to the random sampling principle. Three related objections are encountered. The first is that judgment sampling introduces an undesirable subjectivity into the estimation process. It does have a subjective aspect, to be sure, for when drawing such a sample, a view needs to be taken on which individual characteristics are correlated with the population parameter whose value is being sought, and which are not: a judgment must be made as to whether a person's social class, age, gender, the condition of his front garden, the age of her cat, and so forth, are relevant factors for the sampling process. Without exhaustively surveying the population, you could not pronounce categorically on the relevance of the innumerable, possibly relevant factors; there is, therefore, considerable room for opinions to vary from one experimenter to another. This may be contrasted with random sampling, which requires no individual judgment and is quite impersonal and objective.

The second objection, which is, in truth, an aspect of the first, is that judgment samples are susceptible to bias, due to the experimenter's ignorance, or through the exercise of unconscious, or even conscious, personal prejudices. Yates (1981, pp. 11–16) illustrates this danger with a number of cases where the experimenter's careful efforts to select representative samples were frustrated by a failure to appreciate and take into account crucial variables. Such cases are often held up as a warning against the bias that can intrude into judgment sampling.

Sampling by means of a physical randomizing process, on the other hand, cannot be affected by a selector's partiality or lack of knowledge. On the other hand, it might, by chance, throw up samples that are as unrepresentative as any that could result from the most ill-informed judgment sampling. This seeming paradox[17] is typically turned into a principal advantage in the standard classical response, which says that when sampling is random, the probabilities of different possible samples can be accurately computed and then systematically incorporated into the inference process, using classical estimation methods. But judgment sampling—so the third objection goes—does not lend itself to objective methods of estimation.

There is another strand to the classical response, which invokes the idea of *stratified random sampling*. This involves partitioning

[17] See below, 8.d, for a discussion of Stuart's description of this situation as the "paradox of sampling".

the population into separate groups, or strata, and then sampling at random from each. The classically approved estimate of the population parameter is then the weighted average of the corresponding strata estimates, the weighting coefficients being proportional to the relative sizes of the population and the strata. Stratified random sampling seems clearly intended as a way of reducing the chance of obtaining seriously unrepresentative samples. But the orthodox classical rationale refers instead to the greater 'efficiency' (in the sense defined above) of estimates derived from stratified samples. For, provided the strata are more homogeneous than the population, and significantly different from one another in relation to the quantity being measured, estimation is more 'efficient' using stratified random sampling than ordinary random sampling, and so, by classical standards, it is better.

This rationale is however questionable; indeed, it seems quite wrong. The efficiency of an estimator, it will be recalled, is a measure of its variance. And the more efficient an estimator, the narrower any confidence interval based on it. So, for instance, a stratified random sample might deliver the 95 percent confidence interval 4' 8" ± 3" as an estimate of the average height of pupils in some school, while the corresponding interval derived from an unstratified random sample (which, by chance, is heavily biased towards younger children) might be, say, 3' 2" ± 6". Classical statisticians seem committed to saying that the first estimate is the better one because, being based on a more efficient estimating method, its interval width is narrower. But this surely misappraises the situation. The fact is that the first estimate is probably right and the second almost certainly wrong, but these are words that should not cross the lips of a classical statistician.

Some Advantages of Judgment Sampling

Judgment sampling has certain practical advantages. A pre-election opinion poll, for example, needs to be conducted quickly, and this is feasible with judgment sampling; on the other hand, drawing up a random sample of the population, finding the people who were selected and then persuading them to be interviewed is costly, time consuming, and sometimes impossible, and the election might well be over before the poll has begun. Practical considerations such as these have established the dominance of quota sampling in market research. "Probably 90 percent of all market research uses quota sampling and in most circumstances it is sufficiently reliable to provide consistent results" (Downham 1988, p. 13).

A second point in favour of judgment and quota samples is that they are evidently successful in practice. Opinion polls conducted by their means, insofar as they can be checked against the results of ensuing elections, are mostly more or less accurate, and market research firms thrive, their services valued by manufacturers, who have a commercial interest in accurately gauging consumers' tastes.

A third practical point is that inferences based on non-random samples are often confidently made and believed by others; indeed they seem inevitable when conclusions obtained in one sphere need to be applied to another, as commonly happens. For instance, in a study by Peto *et al.* (1988), a large group of physicians who had regularly taken aspirin and another group who had not showed similar frequencies of heart attacks over a longish period. Upon this basis, the authors of the study advised against adopting aspirin generally as a prophylactic, their implicit and plausible assumption being that the doctors taking part in the study typified the wider population in their cardiac responses to aspirin. Although the rest of the statistical procedures employed in the study were orthodox, this assumption was not checked by means of random samples taken from the population.[18]

We consider the question of sampling methods again when we discuss Bayesian inference in Chapter 8.

5.h Conclusion

Classical estimation theory and significance tests, in their various forms, are still immensely influential; they are advocated in hundreds of books that are recommended texts in thousands of institutions of higher education, and required reading for hundreds of

[18] Smith 1983 makes the same point in relation to another study. On the arguments concerning sampling in this section, see also Urbach 1989.

thousands of students. And the classical jargon of 'statistical significance', 'confidence', and so on, litters the academic journals and has slipped easily into the educated vernacular. Yet, as we have shown, classical 'estimates' are not estimates in any normal or scientific sense, and, like judgments of 'significance' and 'non-significance', they carry no inductive meaning at all. Therefore, they cannot be used to arbitrate between rival theories or to determine practical policy.

A number of other objections that we have explored in this chapter show, moreover, that classical methods are set altogether on the wrong lines, and are based on ideas inimical to scientific method. Principal here is the objection that all classical methods involve an outcome space and hence a stopping rule, which we have argued brings to bear on scientific judgment considerations that are highly counter-intuitive and inappropriate in that context. And classical methods necessarily introduce arbitrary elements that are at variance not only with scientific practice and intuition, but also with the objectivist ideals that motivated them. The founders of the classical philosophy were seeking an alternative to the Bayesian philosophy, which they dismissed as unsuited to inductive method because it was tainted by subjectivity. It is therefore particularly curious and telling that classical methods cannot operate except with their own, hefty subjective input. This was frankly confessed, in retrospect, by one of the founders of the classical approach:

> Of necessity, as it seemed to us [him and Neyman], we left in our mathematical model a gap for the exercise of a more intuitive process of personal judgement in such matters . . . as the choice of the most likely class of admissible hypotheses, the appropriate significance level, the magnitude of worthwhile effects and the balance of utilities. (Pearson 1966, p. 277)

Classical distaste for the subjective element in Bayesian inference puts one in mind of those who were once accused of taking infinite trouble to strain out a gnat, while cheerfully swallowing a camel!

CHAPTER 6
Statistical Inference in Practice: Clinical Trials

We have thus far discussed various methodologies in terms sufficiently abstract to have perhaps created the impression that the question as to which of them is philosophically correct has little practical bearing. But this is far from being the case. We illustrate this point in the present chapter by looking at Classical and Bayesian approaches to the scientific investigation of causal connections, particularly in agricultural (or 'field') and medical (or 'clinical') trials. Large numbers of such trials are under way at any one time; they are immensely expensive; and their results may exert profound effects on farming and clinical practice. And a further practical effect, in the case of clinical trials, is the inconvenience to which the participants are put and the risks to which they may be exposed.

6.a Clinical Trials: The Central Problem

A clinical trial is designed with a view to discovering whether and to what extent a particular drug or medical procedure alleviates certain symptoms, or causes adverse side effects. And a typical goal of an agricultural field trial would be to investigate whether a putative fertilizer increases the yield of a certain crop, or whether a new, genetically engineered potato has improved growth qualities.

Clinical trials typically involve two groups of subjects, all of whom are currently suffering from a particular medical condition; one of the groups, the *test group*, is administered the experimental therapy, while the other, the *control group*, is not; the progress of each group is then monitored over a period. An agricultural trial to compare a new variety of potato (A) with an established

variety (B) might be conducted in a field that is divided into 'blocks', and then subdivided into plots, in which a seed of each variety is sown. The field might look like this:

	Plot 1	*Plot 2*
Block 1	*A*	*B*
Block 2	*A*	*B*
Block 3	*A*	*B*
Block 4	*A*	*B*

The question of causality that is posed in such trials presents a special difficulty, for in order to demonstrate, for example, that a particular treatment cures a particular disease, you need to know not only that people have recovered after receiving the treatment but also that those self-same people would not have recovered if they hadn't received it. This seems to suggest that to establish a causal link you must examine the results of simultaneously treating and not treating the same patients, under identical circumstances, something that is obviously impossible.

An alternative to this impossible ideal might be to conduct the trial with groups of patients who are identical, not in *every* respect, but just in those respects that are causally relevant to the progress of the medical condition under study. Such causally relevant influences are known as *prognostic factors*. Then, if a test group and a control group were properly matched on all the prognostic factors (with the possible exception of the experimental treatment itself), and at the end of the trial they exhibited unequal recovery rates or differed in some other measure of the symptoms, then these variations can clearly be attributed to the experimental treatment. A similar type of inference would also be available, *mutatis mutandis*, in an agricultural trial, provided the seeds and the plants into which they develop were exposed to the same growth-relevant environments. This sort of inference, in which every potential causal factor is laid out and all but one excluded by the experimental information, is a form of what is traditionally called *eliminative induction*.

But the conditions for such an induction cannot be straightforwardly set up. For if you wished to ensure that every prognostic factor will be equally represented in the experimental groups of a trial, you apparently need a comprehensive list of those factors. And as Fisher (1947, p. 18) pointed out, in every situation, there are innumerably many possible prognostic factors, most of which have not even been thought of, let alone tested for relevance; and some of those that have been so tested might have been mistakenly dismissed as causally inactive.

6.b | Control and Randomization

To meet this difficulty, Fisher distinguished between factors that are known to affect the course of the disease, or the growth of the crop, and factors whose influence is unknown and unsuspected. And he claimed that the misleading effects on the inference process of these two kinds of potentially interfering influences could be neutralised by the techniques of *control* and *randomization*, respectively.

Control

A prognostic factor has been 'controlled for' in a trial when it is distributed in equal measure in both the test and the comparison situations. So, for example, a clinical trial involving a disease whose progress or intensity is known to depend on the patient's age would be controlled for that prognostic factor when the test and control groups have similar age structures. Of particular concern in clinical trials is a prognostic factor called the *placebo effect*. This is a beneficial, psychosomatic effect that arises simply from the reassuring feel and encouraging expectations that are created by all the paraphernalia of white-coated, medical attention.[1] The effect is controlled for in clinical trials by creating the

[1] Wall (1999) presents many fascinating examples of placebo effects and argues convincingly that these operate through the subjects' expectations of improvement.

same expectations for recovery in both groups of patients, by treating them in superficially similar ways. Where two different treatments are to be compared, they should be disguised in such a way that the participants in the trial have no idea which they are receiving. And when there is no comparison treatment, the control group would receive a *placebo*, that is to say, a substance that is pharmacologically inert, yet looks and seems exactly like the test drug.[2]

Agricultural trials are sensitive to variations in soil quality and growing conditions. Fisher (1947, p. 64) advised controlling for these factors by planting the seeds in soil "that appears to be uniform, as judged by [its] surface and texture . . . or by the appearance of a previous crop". And the plots on which the planting takes place should be compact, in view of "the widely verified fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are further apart". And when seeds are sown in plant-pots, the soil should be thoroughly mixed before it is distributed to the pots, the watering should be uniform and the developing plants should be exposed to the same amount of light (*ibid.*, p. 41).

Randomization

So much for the known prognostic factors. What of the unknown ones, the so-called 'nuisance variables', whose influences on the course of an experiment have not been controlled for? These, Fisher said, can only be dealt with through 'randomization'. A randomized trial is one in which the experimental units are assigned at random to the trial treatments. So, for example, in a randomized agricultural trial designed to compare the performance of two types of potato, the plots on which each is grown are selected at random. And when testing the efficacy of a drug in a randomized clinical trial, patients are allocated randomly to the test and control groups.

[2] In trials where the test treatment is not a drug, but some other sort of medical or psychological intervention, the placebo takes a correspondingly different form.

This is not the place to discuss the thorny question of what randomness really is. Suffice it to say that advocates of randomization in trials regard certain repeatable experiments as sources of randomness, for example, throwing a die or flipping a coin, drawing a card blindly from a well-shuffled pack, or observing the decay or non-decay in a given time-interval of a radioactive element. Random number tables, which are constructed by means of such random processes, are also regarded as providing effective ways to perform the randomization. If, on the other hand, experimenters formed the experimental groups by allocating patients, say, as the spirit moved them, the randomization rule would not have been respected, for the allocation process would then not have been objectively random in the sense just indicated.

Fisher's randomization rule is widely viewed as a brilliant solution to the problem of nuisance variables in experimental design. In the opinion of Kendall and Stuart (1983, pp. 120–21), it is "the most important and the most influential of [his] many achievements in statistics". Certainly it has been influential, so much so, that clinical and agricultural trials that are not properly randomized are frequently written off as fatally flawed, and regulatory bodies will usually refuse to license drugs that have not been tested in randomized trials.

We need to examine the far-reaching claims that are made for the virtues of randomization, particularly in view of the fact, which we shall shortly explain, that applying it may be costly, inconvenient and ethically challenging. We shall in fact argue that the standard defences of the procedure as an absolute necessity are unsuccessful and that the problem of nuisance variables in trials cannot be solved by its means. In Chapter 8, we argue that although randomization may sometimes be harmless and even helpful, it is not a *sine qua non*, not absolutely essential. We shall maintain that the essential feature of a trial that permits a satisfactory conclusion as to the causal efficacy of a treatment is the presence of adequate controls.

Two main arguments for randomization are propounded in the literature. The first, which is based on the classical statistical idea that significance tests are the central inferential tool, claims that such tests are valid only if the experiment was randomized. The second argument, while also advanced by those of

a classical outlook, evidently appeals to a modified form of eliminative induction.

6.c | Significance-Test Defences of Randomization

Fisher was concerned by the fact that the comparison groups in a clinical trial might be badly mismatched through the unequal distribution of one or more of the infinitely many possible, unknown nuisance variables. For, in the event of such a mismatch, you might be badly misled if, for example, you concluded definitely that the group having the greater recovery rate had received the superior treatment. Corresponding concerns arise with field trials.

Fisher's response was to say that although we cannot know for certain that the groups are perfectly matched in regard to every prognostic factor, the randomization process furnishes us with objective and certain knowledge of the probabilities of different possible mismatches, knowledge that is required for a valid test of significance. And of course, Fisher favoured such tests as a way of drawing conclusions from experimental results.

> [T]he full procedure of randomisation [is the method] by which the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated [by being controlled]. (Fisher 1947, p. 19)

Fisher's reason for regarding randomization as a necessary feature of clinical trials has been widely accepted amongst classically minded statisticians. For example, Byar *et al.* (1976, p. 75): "It is the process of randomization that generates the significance test, and this process is independent of prognostic factors, *known or unknown*" (our italics). This argument will not cut the mustard with Bayesians, who do not acknowledge significance tests as valid forms of inductive reasoning. But as we shall argue, it is incorrect even in the classical context.

The Problem of the Reference Population

Fisher claimed that significance tests could only be applied to clinical and agricultural trials if these were randomized. It is

therefore surprising that expositions of the standard significance tests employed in the analysis of trial results, such as the *t*-test, the chi-square test, and the Wilcoxon Rank Sum test, barely allude to the proviso. Take the case of the *t*-test applied to a clinical trial whose goal was to determine whether or not a particular treatment has a beneficial effect on some medical disorder. In the trial, a certain 'population' of people who are suffering from the condition in question is divided between a test and a control group by the classically required randomization method. Suppose that the trial records some quantitative measure of the disease symptoms, giving means of $\bar{x}_1$ and $\bar{x}_2$, respectively, in the test and control groups. The difference $\bar{x}_1 - \bar{x}_2$ is likely to vary from experiment to experiment, and the associated probability distribution has a certain standard deviation, or 'standard error', that is given by

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

where σ_1 and σ_2 are the population standard deviations,[3] and where n_1 and n_2 are, respectively, the sizes of two samples. If the null hypothesis affirms that the treatment has no average effect, a test-statistic that may be employed in a test of significance is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}.$$

The situation is complicated because the two standard deviations are usually unknown, in which case it is recommended that they be estimated from the standard deviations in the corresponding samples. And provided certain further conditions are met, the statistic that is then proposed for use in the test of significance of the null hypothesis is the *t*-statistic obtained from Z by substituting the corresponding sample standard deviations for σ_1 and σ_2.

This and similar significance tests, as well as related estimating procedures, are directed to detecting and estimating differences between population parameters, and their validity therefore

[3] These are hypothetical standard deviations. That is, σ_1 (respectively, σ_2) is the standard deviation that the population would have if all its members were treated with the test treatment (respectively, the placebo).

depends on the experimental samples being drawn at random from the population in question. But which population is that?

The randomization step provides a possible answer, and a possible explanation for why it is involved in the testing process. For, by randomizing the subjects across the experimental groups, each of the groups is certain to be a random selection from the subjects participating in the trial. This explanation implies that the reference population comprises just those patients who are involved in the trial. But this will not do, for the aim of a trial is to determine not just whether the treatment was effective for those who happen to be included in the trial but whether it will benefit others who were left out, as well as unknown sufferers in faraway places, and people presently healthy or yet unborn who will contract the disease in the future. But in no trial can random samples be drawn from hypothetical populations of notional people.

Bourke, Daly, and McGilvray (1985, p. 188) addressed the question of the reference population slightly differently but they too saw it resolved through the randomization step. They argued that the standard statistical tests may be properly applied to clinical trials without "the need for convoluted arguments concerning random sampling from larger populations", provided the experimental groups were randomized. To illustrate their point they considered a case where the random allocation of 25 subjects delivered 12 to one group and 13 to the other. These particular groups, they affirmed, "can be viewed as one of many possible allocations resulting from the randomization of 25 individuals into two groups of 12 and 13, respectively. They noted the existence of 5,200,300 different possible outcomes of such a randomization and proposed this set of possible allocations as the reference population; the null hypothesis to which the significance test is applied, they suggested, should then be regarded as referring to that hypothetical population.

But this suggestion is not helpful. First, it is premised on the fiction of a premeditated plan only to accept groups resulting from the randomization that contain, respectively, 12 and 13 subjects, with the implication that any other combinations that might have arisen would have been rejected.[4] Secondly, while the reference population consists of five million or so pairs of groups, the sample is just a single pair. But as is agreed all round, you get very little information about the mean of a large population from a sample of 1. Finally, as Bourke *et al.* themselves admitted, they do not overcome the problem we identified above, for statistical inference in their scheme "relates only to the individuals entered into the study" and may not generalize to any broader category of people. Achieving such a generalization, they argued, "involves issues relating to the representativeness of the trial group to the general body of patients affected with the particular disease being studied". But the path from "representative sample" to "general body of patients"—two vague notions—cannot be explored via significance tests and is left uncharted; yet unless that path is mapped out, the randomized clinical trial can have nothing at all to say on the central issue that it is meant to address.

[4] The numbers of subjects, n_1 and n_2, in each of the groups must, strictly, be fixed in advance for a valid t-test, for reasons we discussed earlier (5.d). One way of combining this and the randomization requirement would be to select pairs of subjects at random and then allocate the elements of the pairs at random, one to the test and the other to the control group. When the number of subjects is even, n_1 and n_2 will be equal, and when odd, they will differ by one. This may be what Bourke, Daly, and McGilvray had in mind.

Fisher's Argument

When Fisher first advanced the case for randomization as a necessary ingredient of trials he did so in somewhat different terms. Instead of trying to relate the significance test to populations of sufferers, he thought of it as testing certain innate qualities of a medical treatment or, in the agricultural context, where Fisher's main interest lay, innate potentials of a plant variety.

We may illustrate the argument through our potato example, which slightly simplifies Fisher's (1947, pp. 41–42) own example. Consider the possibility that, in reality, the varieties A and B have the same genetic growth characteristics and let this be designated the null hypothesis. Suppose, too, that one plot in each block is more fertile than the other, and that this apart, no relevant difference exists between the conditions that the seeds and the resulting plants experience. If pairs of different-variety seeds are allocated

at random, one to each of a pair of plots, there is a guaranteed, objective probability of exactly a half that any plant of the first variety will exceed one of the second in yield, even if no intrinsic difference exists. The objective probability that r out of n pairs show an excess yield for A can then be computed, a simple test of significance, such as we discussed in Chapter 5, applied, and a conclusion on the acceptability or otherwise of the null hypothesis drawn.

All this depends on the objective probabilities that the randomization step is supposed to guarantee. But this guarantee in turn depends on certain questionable, suppositions, for example, that none of the innumerable environmental variations that emerged *after* the random allocation introduced a corresponding variation in plant growth. For if the experiment had been exposed to some hidden growth factor, the null-hypothesis probabilities might be quite different from those assumed above, nor could we know what they are.

Fisher (1947, p. 20) recognized this as a difficulty, but argued that it could be overcome: "the random choice of the objects to be treated in different ways would be a *complete guarantee* of the validity of the test of significance, if these treatments were the last in time of the stages in the physical history of the objects which might affect their experimental reaction" (italics added). And any variation that occurred after this, carefully located, randomization step, he then claimed, "causes no practical inconvenience" (1947, pp. 20–21),

> for subsequent causes of differentiation, if under the experimenter's control . . . can either be predetermined before the treatments have been randomised, or, if this has not been done, can be randomised on their own account: and other causes of differentiation will be either (a) consequences of differences already randomised, or (b) natural consequences of the differences in treatment to be tested, of which on the null hypothesis there will be none, by definition, or (c) effects supervening by chance independently from the treatments applied.

In other words, no post-randomization effects could disturb the probability calculations needed for the significance test, for such effects would either be the product of differences already randomized and hence would be automatically distributed at random, or

they would be chance factors, independent of the treatment, and therefore subject to a spontaneous randomization.

But Fisher is here neglecting the possibility of influences (operating either before or after the random allocation) that are *not* independent of the treatments. For example, a certain insect might benefit plant varieties to an equal degree when these are growing separately, but be preferentially attracted to one of them when they are in the same vicinity. Or the different varieties in the trial might compete unequally for soil nutrients. There are also possible disturbing influences that could come into play before the plants were sown. For instance, the different types of seed might have been handled by different market gardeners, or stored in slightly different conditions, and so on. Factors such as these might, unknown to the experimenters, impart an unfair advantage to one of the varieties, and their nature is such that they cannot be distributed at random, either spontaneously or through the trial's deliberate randomization. Fisher (1947, p. 43) was therefore wrong in his view that randomization "relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which the data may be disturbed."

The natural response to this objection is to say that while one or more of the innumerable variations occurring during an experiment might possibly affect the result, most are very unlikely to do so. For example, the market gardener's hair colour and the doctor's collar size are conceivable, but most implausible as influences on the progress of an agricultural or clinical trial. It seems reasonable therefore to ignore factors like these in the experimental design, and this is also the official position:

> A substantial part of the skill of the experimenter lies in the choice of factors to be randomised out of the experiment. If he is careful, he will randomise out [distribute at random] all the factors which are suspected to be causally important but which are not actually part of the experimental structure. But every experimenter necessarily neglects some conceivably causal factors; if this were not so, the randomisation procedure would be impossibly complicated. (Kendall and Stuart 1983, p. 137)

In accordance with this doctrine, when designing a trial to study the effect of alcohol on reaction times, Kendall and Stuart

explicitly omitted the subjects' eye colour from any randomization, on the grounds that the influence of this quality was "almost certainly negligible". They cited no experimental results in support of their view, presumably because there are none. But even if there had been a trial to which they could have appealed, certain conceivable influences on its outcome must also have been set aside as negligible. Hence, it would be futile to insist that an influence must only be considered negligible in the light of a properly randomized trial designed to determine this, for that would open up the need for an infinite regress of trials. This is why Kendall and Stuart (p. 137) concluded that the decision whether to "randomize out a factor" or ignore it "is essentially a matter of judgement".[5]

What Kendall and Stuart demonstrated is that randomization has to be confined to factors that the experimental designers judge to be of importance, and that this judgment is necessarily a personal one, which cannot be based solely on objective considerations. This conclusion is of course completely at odds with the classical methodology underlying Fisher's argument. So, far from rescuing Fisher's defence of randomization, Kendall and Stuart unwittingly knocked another nail in its coffin.

6.d The Eliminative-Induction Defence of Randomization

This defence, at its strongest, claims that randomization performs for the unknown prognostic factors the same service as controls perform for the known ones; that is to say, the procedure guarantees that both the known *and* the unknown prognostic factors in a trial will be equally distributed across the experimental groups. If

[5] It should be mentioned in passing that Kendall and Stuart's example is not well chosen, for in their experiment, the different doses of alcohol are allocated randomly to the subjects, and this means that any characteristic to which the subjects are permanently attached, like the colour of their eyes, is, contrary to their claim, automatically randomized. But their point stands, for there are other examples they could have used—the eye colour of those who administer the alcohol to the subjects is one.

such an assurance were available, it would then be a straightforward matter to infer causal responsibility for any discrepancies that arise in the course of the trial between the experimental groups.

These are some expressions of the defence: "Allocating patients to treatments *A* and *B* by randomization produces two groups of patients which are as alike as possible with respect to *all* their [prognostic] characteristics, *both known and unknown*" (Schwartz *et al.*, 1980, p. 7; italics added). Or as Giere (1979, p. 296) emphatically put it: randomized groups "are automatically controlled for ALL other factors, even those no one suspects."

The argument, however, usually takes a more modest form. For example, Byar *et al.* (1976, pp. 74–80; our italics) say that randomization "*tends to* balance treatment groups in . . . prognostic factors, whether or not these variables are known." Tanur *et al.* (1989, p. 10; our italics) maintain that it "*usually* will do a good job of evening out all the variables—those we didn't recognize in advance as well as those we did recognize." And Gore (1981, p. 1559; our italics) expresses the idea more precisely by saying that randomization is "an insurance *in the long run* against substantial accidental bias between treatment groups", indicating that "the long run" covers large trials with more than 200 subjects. In other words, in the view of these commentators, although randomization does not give a complete assurance that the experimental groups will be balanced, it makes such an outcome very probable, and the larger the groups, the greater that probability.

This latter claim is certainly credible, judging by the frequency with which it is voiced. Nevertheless, we argue that it is mistaken, unless significantly modified. And this modified position, while compatible with Bayesian thought, is inimical to classical inferential theory.

Bearing in mind that the argument under examination makes claims about the *unknown* prognostic factors operating in a trial, we may consider a number of possibilities regarding their number and nature. The simplest possibility is that there are no such factors, in which case, the probability that the trial throws up unmatched groups is clearly zero. The next possibility is that there is a single unknown factor—call it *X*—that is carried by some of the patients. The randomized groups will be substantially

unmatched on X with a certain definite probability, x_n, say, and this probability clearly diminishes as the sample size, n, increases.

There might, of course, be a second factor, Y, also carried by some of the patients. The chance of one of the groups created in the randomization step containing substantially more Y-patients than the other also has a definite value, y_n, say. And the probability that the groups are unmatched on at least one of the two factors might then be as much as $x_n + y_n - x_n y_n$, if they are independent. And because we are dealing with the unknown, it must be acknowledged that there might be innumerable other factors; hence the probability of a substantial imbalance on *some* prognostic factor might, for all we know, be quite large, as Lindley (1982, p. 439) has pointed out. Nor are we ever in a position to calculate how large, since, self-evidently, we do not know what the unknown factors are.

We have so far considered only unknown factors that are, as it were, attached to patients. What about those that might be independent of the patient but connected to the treatment? We are here thinking of possible, unknown prognostic factors that are accidentally linked to the treatment in this particular trial, so that they could not be regarded as an aspect of the treatment itself. For example, suppose the test and control patients were treated in separate surgeries whose different environments, through some hidden process, either promoted or hindered recovery; or imagine that the drug, despite the manufacturers' best efforts, included a contaminant which compromised its effectiveness; or . . . (one simply needs a rich enough imagination to extend this list indefinitely). For all we know, one or more such factors are active in the experimental situation; and if that were in fact so, the probability that the groups are imbalanced would, of course, be one.

So what can we say with assurance about the objective probability of the randomized experimental groups in a clinical trial differing substantially on unknown prognostic factors? Merely this: it lies somewhere in the range zero to one! It follows that the main defence of randomization in trials—that it guarantees well-matched groups, or ensures that such groups are highly probable—is untrue.

Those who advance either of the claims that we have just shown to be faulty do so in the belief that drawing conclusions

from the results of a clinical trial would be facilitated by a guarantee that the comparison groups are balanced, or very probably balanced on every prognostic factor. Take the stronger of these claims, which is in fact rarely maintained and clearly indefensible. It has at least the virtue that if it were true, then the conditions for an eliminative induction would be met, so that whatever differences arose between the groups in the clinical trial could be infallibly attributed to the trial treatment. On the other hand, the less obviously faulty and more popular claim cannot exploit eliminative induction. For, the premise that the experimental groups were *probably* balanced does not imply that differences that arise in the clinical trial were *probably* due to the experimental treatment, unless Bayes's theorem were brought to bear, but that would require the input of prior probabilities and the abandonment of the classical approach.

We shall, in Chapter 8, see how Bayesian inference operates in clinical research, in the course of which we shall show that randomization, while it may sometimes be valuable, is not absolutely necessarily.

6.e | Sequential Clinical Trials

There are two aspects of the controlled, randomized trials we have been considering that have caused concern, even amongst classical statisticians. The first is this. Suppose that the test therapy and the comparison therapy (which may simply be a placebo) are not equally effective and that this becomes apparent at the end of the trial, through a differential response rate in the experimental groups. This would mean that some of the patients had been treated with an inferior therapy throughout the trial. Clearly the fewer so treated the better. So the question arises whether a particular trial is suited to extracting the required information in the most efficient way.

The second concern is prompted by the following consideration: as we showed in an earlier chapter, any significance test that is applied to the results of a trial requires the experimenter to establish a stopping rule in advance. The stopping rule that is normally used, or assumed to have been used, fixes the number of

patients who will enter the trial. Then, if for some reason, the trial were discontinued before the predetermined number was reached, the results obtained at that stage would, as Jennison and Turnbull (1990, p. 305) note, have to be reckoned quite uninformative. Intuitively, however, this is wrong; in many circumstances, quite a lot of information might seem to be contained in the interim results of a trial. Suppose, to take an extreme example, that a sample of 100 had been proposed and that the first 15 patients in the test group completely recover, while a similar number in the control group promptly experience a severe relapse. Even though the trial as originally envisaged is incomplete, this evidence would incline most people strongly in favour of the test treatment and would make them reluctant to continue the trial, fearing that a further 35 or so patients would be denied the apparently superior treatment.

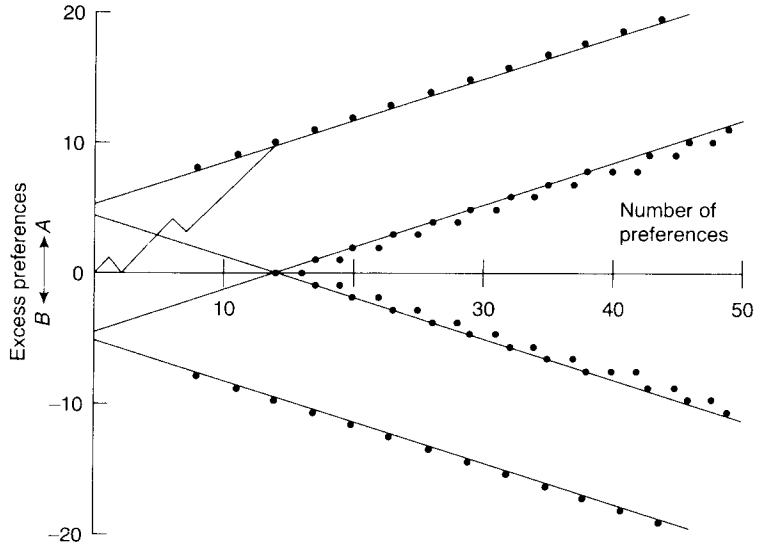
It might be imagined that the problem could be dealt with by performing a new significance test after each new experimental reading and then halting the trial as soon as a 'significant' result is obtained. But, in fact, this sort of continuous assessment is not possible, for each significance test would be based on the assumption that the number of patients treated at that stage of the trial was the number envisaged in the stopping rule for the entire trial, which of course is not the case. A kind of clinical trial in which sequences of such quasi-significance tests may be validly used has, however, been developed and we examine these now.

Sequential clinical trials were designed to minimize the number of patients that need to be treated and to ensure that "random allocation should cease if it becomes reasonably clear that one treatment has a better therapeutic response than another" (Armitage 1975, p. 27). Such trials enable tests of significance to be applied even when the sample size is variable; they allow results to be monitored as they emerge; and in many cases, they achieve a conclusion with fewer data than a fixed-sample design would allow.

The following is a simple sequential trial, which we have taken from Armitage 1975. Appropriately matched pairs of patients are treated in sequence either with substance A or substance B (one of which might be a placebo). The treatments are randomized, in the sense that the element of any pair that receives

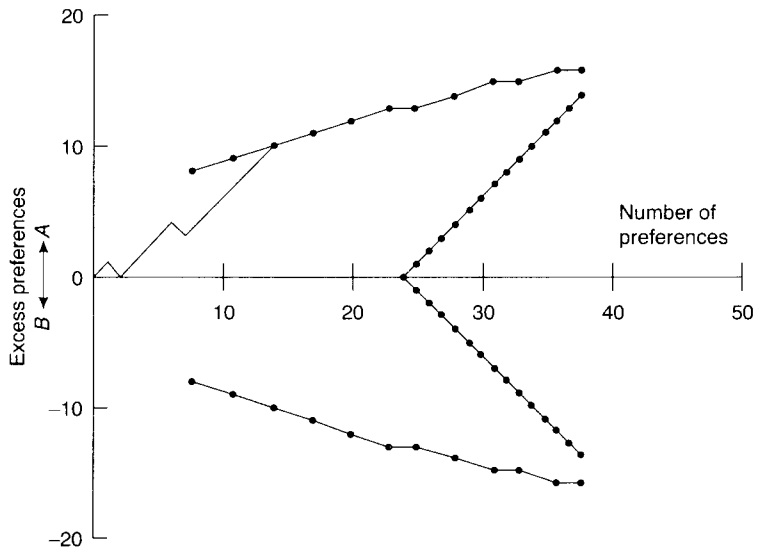
A, respectively B, is determined at random. We shall assume that A and B are indistinguishable to the patients and that the medical personnel administering them are also in the dark as to how each patient was treated (the trial is therefore both 'blind' and 'double-blind'). For each pair, a record is made of whether, relative to some agreed criterion, the A- or the B-treated patient did better. The results are assumed to come in sequentially and to be monitored continuously.

Results are recorded for each pair as either A doing better than B, or B doing better than A; we label these results simply **A** and **B**, respectively. In the diagrams below, the difference between the number of **A**s and the number of **B**s in a sample is plotted on the vertical axis; the horizontal axis represents the number of comparisons made. The zigzag line, or 'sample path', starting from the origin, represents a possible outcome of the trial, corresponding to the 14 comparisons, **ABAAAABAAAAAAA**.



SEQUENTIAL PLAN *a*

The wedge-shaped system of lines in sequential plan a was designed to have the following characteristics: if the drugs are equally effective (the null hypothesis), then the probability of the sample path first crossing one of the outer lines is 0.05. If the



SEQUENTIAL PLAN *b*

sample path does cross one of these lines, sampling is stopped and the null hypothesis rejected. If the sample path first crosses one of the inner lines, the sampling is stopped too, the result declared non-significant, and the null hypothesis accepted. The inner lines are constructed, in this example, so that the test has a power of 0.95 against a particular alternative hypothesis, namely that the true probability of an *A*-treated patient doing better than a *B*-treated patient is 0.8.

The sequential test based on plan *a* is, in Armitage's terminology, 'open', in the sense that the trial might never end, for the sample path could be confined indefinitely within one of the two channels formed by the two pairs of parallel lines. Plan *b* represents a 'closed' test of the same null and alternative hypotheses. In the latter test, if the sample path meets one of the outer lines first, the null hypothesis is rejected, and if one of the inner lines, the hypothesis is accepted. This closed test has a similar significance level and power to the earlier, open one, but differs from it in that at most 40 comparisons are required for a decision to be reached.

The above sequential plans were calculated by Armitage on the assumption that as each new datum is registered, a new significance test, with an appropriate significance level, is carried out. (As we pointed out, these are quasi-significance tests, since they presume a fixed sample size when, in fact, the sample is growing at every stage.) The significance level of the individual tests is set at a value that makes the overall probability of rejecting a true null hypothesis 5 percent. This is just one way of determining sequential tests, particularly favoured by Armitage (1975), but there are many other possible methods leading to different sequential plans. The sequential tests might also vary in the frequency with which data are monitored; the results might be followed continuously as they accumulate or examined after every new batch of, say, 5 or 10 or 50 patients have been treated (these are 'group sequential trials'). Each such interim-analysis policy corresponds to a different sequential test.

The existence of a multiplicity of tests for analysing trial results carries a strange, though by now familiar implication: whether a particular result is 'significant' or not, whether a hypothesis should be rejected or not, whether or not you would be well advised to take the medicine in the expectation of a cure, depend on how frequently the experimenter looked at the data as they accumulated and which sequential plan he or she followed. Thus a given outcome from treating a particular number of patients might be significant if the experimenter had monitored the results continuously, but not significant if he had waited to analyse the results until they were all in, and the conclusion might be different again if another sequential plan had been used. This is surely unacceptable, and it is no wonder that sequential trials have been condemned as "a hoax" (Anscombe 1963, p. 381). As Meier (1975, p.525) has put it, "it seems hard indeed to accept the notion that *I* should be influenced in my judgement by how frequently *he* peeked at the data while he was collecting it."

Nevertheless, the astonishing idea that the frequency with which the data have been peeked at provides information on the effectiveness of a medical treatment and should be taken account of when evaluating such a treatment is thoroughly entrenched. Thus the Food and Drug Administration, the drugs licensing authority in the United States, includes in its *Guideline* (1988, p. 64) the requirement that "all interim analyses, formal or informal, by any study participant, sponsor staff member, or data monitor-

ing group should be described in full . . . . The need for statistical adjustment because of such analyses should be addressed. Minutes of meetings of the data monitoring group may be useful (and may be requested by the review division)". But the intentions and calculations made by the various study participants during the course of a trial are nothing more than disturbances in those people's brains. Such brain disturbances seem to us to carry as much information about the causal powers of the drug in question as the goings-on in any other sections of their anatomies: none.

6.f | Practical and Ethical Considerations

The principal point at issue in this chapter has been Fisher's thesis, so widely adopted as an article of faith, that clinical and agricultural trials are worthless if they are not randomized. But as we have argued, the alleged absolute need for such trials to be randomized has not been established. In Chapter 8, we shall argue that a satisfactory approach to the design of trials and the analysis of their results is furnished by Bayes's theorem, which does not treat randomization as an absolute necessity.

It would not, of course, be correct to infer that randomization is necessarily harmful, nor that it is never useful—and we would not wish to draw such a conclusion—but removing the absolute requirement for randomization is a significant step which lifts some severe and, in our view, undesirable limitations on acceptable trials. For example, the requirement to randomize excludes as illegitimate trials using so-called historical controls. Historical controls suggest themselves when, for example, one wishes to find out whether a new therapy raises the chance of recovery from a particular disease compared with a well-established treatment. In such cases, since many patients would already have been observed under the old regime, it seems unnecessary and extravagant to submit a new batch to that same treatment. The control group could be formed from past records and the new treatment applied to a fresh set of patients who have been carefully matched with those in the artificially constructed control group (or historical control). But the theory of randomization prohibits this kind of experiment, since patients are not assigned with equal probabilities to the two groups; indeed, subjects finding themselves in either of the groups would have had no chance at all of being chosen for the other.

There is also sometimes an unattractive ethical aspect to randomization, particularly in medical research. A new treatment, which is deemed worth the trouble and expense of an investigation, has often recommended itself in extensive pilot studies and in informal observations as having a reasonable chance of improving the condition of patients and of performing better than established treatments. But if there were evidence that a patient would suffer less with the new therapy than with the old, it would surely be unethical to expose randomly selected sufferers to the established and apparently or probably inferior treatment. Yet this is just what the theory of randomization insists upon. No such ethical problem arises when patients receiving the new treatment are compared with a matched set of patients who have already been treated under the old regime.

6.g | Conclusion

We have reviewed the main features of classically inspired designs for clinical and agricultural trials, particularly the alleged requirement for treatments to be randomized over the experimental units. Our conclusion is that neither of the two standard arguments for the randomization principle is effective and that the principle is indefensible as an absolute precondition on trials, even from the classical viewpoint. In Chapter 8, we shall argue in detail that a Bayesian approach gives a more satisfactory treatment of the problem of nuisance variables and furnishes intuitively correct principles of experimental design.

CHAPTER 7
Regression Analysis

We are often interested in how some variable quantity depends on or is related to other variable quantities. Such relationships are often guessed at from specific values taken by the variables in experiments. As we pointed out earlier (in 4i), infinitely many different, and conflicting relationships are compatible with any set of data. And yet, despite the immense multiplicity of candidate theories or relationships, scientists are rarely left as bewildered as might be imagined; in fact, they often become quite certain about what the true relationships are and feel able to predict hitherto unobserved values of the variables in question with considerable confidence. How they do this and with what justification is the topic of this chapter (which largely follows Urbach 1992).

7.a | Simple Linear Regression

How to infer a general relationship between variable quantities from specific observed values is a problem considered in practically every textbook on statistical inference, under the heading *Regression.* Usually, the problem is introduced in a restricted form, in which just two variables are involved; the methods developed are then extended to deal with many-variable relationships. When it comes to examining underlying principles, which is our aim, the least complicated cases are the best and most revealing; so it is upon these that we shall concentrate.

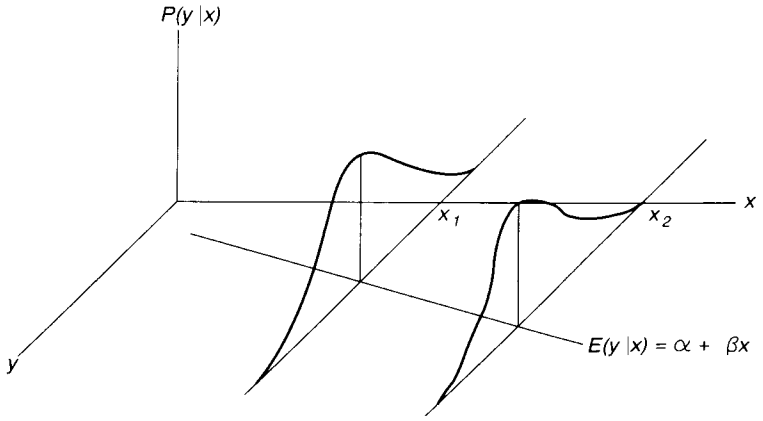
Statisticians treat a wider problem than that alluded to earlier, by allowing for cases where the variables are related, but because of a random 'error' term, not in a directly functional way. With just two variables, x and y, such a relationship, or regression, may be represented as $y = f(x) + \in$, where $\in$ is a random variable,

called the 'error' term. The simplest and most studied regressions—known as *simple linear regressions*—are a special case, in which $y = \alpha + \beta x + \in$, α and β being unspecified constants. In addition, the errors in a simple linear regression are taken to be uncorrelated and to have a zero mean and an unspecified variance, σ^2, whose value is constant over all x. We shall call regressions meeting these conditions linear, dropping the qualification 'simple', for brevity.

Many different systems have been explored as reasonable candidates to satisfy the linear regression hypothesis. The following are typical examples from the textbook literature: the relationship between wheat yield, y, and quantity of fertilizer applied, x (Wonnacott and Wonnacott 1980); the measured boiling point of water, y, and barometric pressure, x (Weisberg 1980); and the level of people's income, y, and the extent of their education, x (Lewis-Beck 1980).

The linear regression hypothesis is depicted below (Mood and Graybill 1963, p. 330). The vertical axis represents the probability densities of y for given values of x. The bell-shaped curves in the y, $P(y \mid x)$-plane illustrate the probability density distributions of y for two particular values of x; the diagonal line in the x,y-plane is the plot of the mean of y against x; and the variance of the density distribution curves, which is the same as the error variance, σ^2, is constant—a condition known as 'homoscedasticity'.

The linear regression hypothesis leaves open the values of the three parameters α, β, and σ, which need to be estimated from



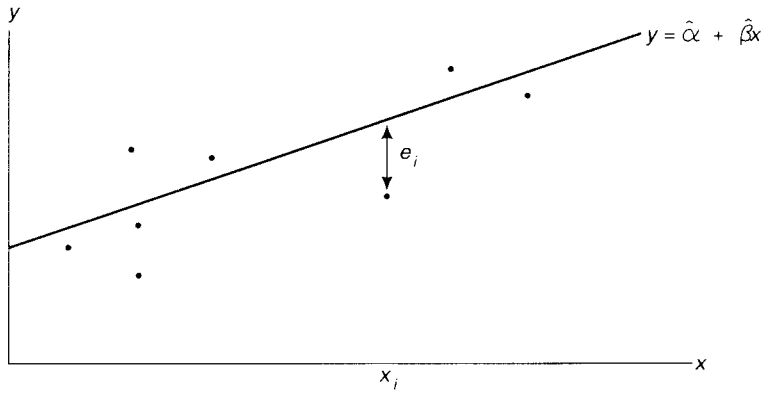
data, such data normally taking the form of particular x, y readings. These readings may be obtained in one of two ways. In the first, particular values of x (such as barometric pressures or concentrations of fertilizer) are pre-selected and then the resulting ys (for instance the boiling point of water at each of the pressure levels or wheat yields at each fertilizer strength) are read off. Alternatively, the xs may be selected at random (they could, for example, be the heights of random members of a population) and the corresponding ys (for example each chosen person's weight) then recorded as before. In the examples discussed here, the data will mostly be of the former kind, with the xs fixed in advance; none of the criticisms we shall offer will be affected by this restriction.

We are considering the special case where the underlying regression has been assumed to take the simple linear form. For the Bayesian, this means that other possible forms of the regression all have probability zero, or their probabilities are sufficiently close to zero to make no difference. In such special (usually artificial) cases, a Bayesian would continue the analysis by describing a prior, subjective probability distribution over the range of possible values of the regression parameters α, β, and σ, and would then conditionalise on the evidence to obtain a corresponding posterior distribution (for a detailed exposition, see, for instance, Broemeling 1985). But classical statisticians regard such distributions as irrelevant to science, because of their subjective component. The classical approach tries to put its objectivist ideal into effect by seeking the linear regression equation possessing what is called the 'best fit' to the data, an aim also expressed as a search for the 'best estimates' of the unknown linear regression parameters. There are also classical procedures for examining the assumption of linear regression, which we shall review in due course.

7.b | The Method of Least Squares

The method of least squares is the standard classical way of estimating the constants in a linear regression equation and "historically has been accepted almost universally as the best estimation technique for linear regression models" (Gunst and Mason 1980, p. 66).

The method is this. Suppose there are n x,y readings, as depicted in the graph below. The vertical distance of the ith point from any straight line is labelled e_i and is termed the 'error' or 'residual' of the point relative to that line. The straight line for which Σe_i^2 is minimal is called the *least squares line.* If the least squares line is $y = \hat{\alpha} + \hat{\beta}x$, then $\hat{\alpha}$ and $\hat{\beta}$ are said to be the least squares estimates of the corresponding linear regression coefficients, α and β, and the line is often said to provide the 'best fit' to the data. The idea is illustrated below.



The least squares line, that is, the line for which Σe_i^2 is minimum.

The least squares estimates, $\hat{\alpha}$ and $\hat{\beta}$, are given as follows, where $\bar{x}$ and $\bar{y}$ are, respectively, the mean values of the x_i and y_i in the sample (proving these formulas is straightforward and we leave this to the reader):

$$\hat{\beta} = \frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

These widely used formulas, it should be noted, make no assumptions about the distribution of the errors. Note, too, that the least squares principle does not apply to σ^2, the error variance, which needs to be estimated in a different way (see below).

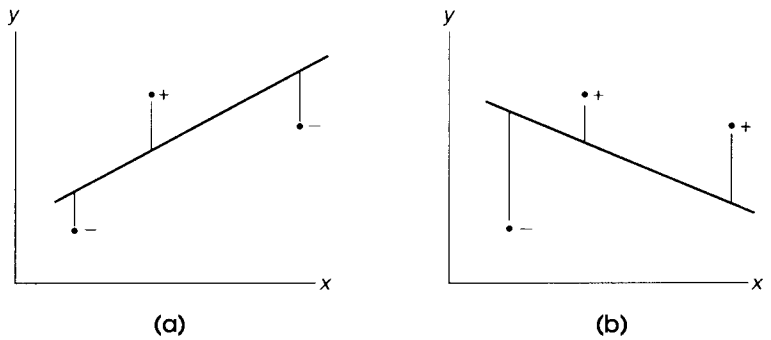
7.c Why Least Squares?

It is said almost universally by classical statisticians that if the regression of y on x is linear, then least squares provides the best estimates of the regression parameters. The term 'best' in this context intimates some epistemic significance and suggests that the least squares method is for good reason preferable to the infinitely many other conceivable methods of estimation. Many statistics textbooks adopt the least squares method uncritically, but where it is defended, the argument takes three forms. First, the method is said to be intuitively correct, and secondly and thirdly, to be justified by the Gauss-Markov theorem and by the Maximum Likelihood Principle. We shall consider these lines of defence in turn.

Intuition as a Justification

The least squares method is often recommended for its "intuitive plausibility" (Belsley *et al.* 1980, p. 1). Yet the intuitions typically cited do not point unequivocally to least squares as the right method for estimating the parameters of a linear regression equation; at best, they are able to exclude some possible alternative methods. For example, the idea that one should minimize the absolute value of the sum of the errors, $|\Sigma e_i|$, is often dismissed because, in certain circumstances, it leads to intuitively unsatisfactory regression lines. In the case illustrated below, the method would result in two, quite different lines for the same set of data. Although both lines minimize $|\Sigma e_i|$, Wonnacott and Wonnacott (p. 16) judge one of them, namely b, to be "intuitively . . . a very bad one".

The trouble here is that the criterion of minimum aggregate error can be met by balancing large positive errors against large negative ones, whereas for an intuitively good fit of data to a line, all the errors should be as small as possible. This difficulty would be partly overcome by another suggested method, namely, that of minimizing the sum of the absolute deviations, $\Sigma|e_i|$. But this too is often regarded as unsatisfactory. For example, Wonnacott and Wonnacott rejected the method unequivocally, and signalled their



Two lines both satisfying the condition that $|\Sigma e_i^2|$ is minimized.

emphatic disapproval of the method by giving it the acronym MAD. But this highlights a weakness of arguments that are based on intuition, namely, that people often disagree over the intuitions themselves. Thus, unlike the Wonnacotts, Brook and Arnold (1985, p. 9) found the MAD method perfectly sane; indeed, they regarded it as "a sensible approach which works well", their sole reservation being the practical one that "the actual mathematics is difficult when the distributions of estimates are sought". (As we shall see in the next section, statisticians require such distributions, in order to derive confidence intervals for predictions based on an estimated regression equation.)

On the other hand, Brook and Arnold saw in the least squares method an intuitively unsatisfactory feature. This is that particularly large deviations, since they must be squared, "may have an unduly large influence on the fitted curve" (1985, p. 9). They refer with approval to a proposal for mitigating this supposed drawback, according to which the estimation should proceed by minimizing the sum of a lower power of the errors than their squares, that is, by minimizing $\Sigma |e_i|^p$, with p lying somewhere between 1 and 2. The authors propose $p = 1.5$ as a "reasonable compromise", but they do not explain why the mid-point value should accord so well with reason; to us it seems merely arbitrary, despite its air of Solomonic wisdom. But against this proposal, they point out the same practical difficulty that they observed with the MAD method of estimation, namely that "it is difficult to determine the exact distributions of the resulting estimates", and so—for this purely pragmatic reason—they revert to $p = 2$.

The deviations whose squares are minimized in the least squares method appear in a graph as the vertical distances of points from a line. It might be thought that a close fit between line and points could be equally well secured by applying the least squares principle to the perpendicular distances of the points from the line, or to the horizontal distances, or, indeed, to distances measured along any other angle. Why should the vertical deviations be privileged above all others?

Kendall and Stuart defended the usual least squares procedure, based on vertical distances, on the vague and inadequate grounds that since "we are considering the dependence of y on x, it *seems natural* to minimize the sum of squared deviations in the y-direction" (1979, p. 278; italics added). But this by itself is insufficient reason, for many procedures that we find natural are wrong; indeed, Kendall and Stuart's famous textbook offers numerous 'corrections' for what they view as misguided intuitions.

Brook and Arnold advanced an apparently more substantial reason for concentrating on vertical deviations from the regression line; they claimed that "when our major concern is predicting y from x, the vertical distances are more relevant *because they represent the predicted error*" (1985, p. 10; italics added). For most statisticians the main interest of regression equations does indeed lie in their ability to predict (see Section 7.e below). But predictions concern previously unexamined values of the variables; the predicting equation, on the other hand, is built up from values that have already been examined. It is misleading therefore to say that "vertical distances [of data points] are more relevant because they represent the predicted error"; in fact, they do not represent the errors in predictions of previously unexamined points.

The following seems a more promising argument for minimizing the squares of the vertical distances. Suppose the units in which y was measured were changed, by applying a linear transformation, $y' = my + n$, m and n being constants. The transformed regression equation would then be $y' = \alpha' + \beta' x + \in$, where $\alpha' = m\alpha + n$ and $\beta' = m\beta$. Ehrenberg (1975) and others (e.g., Bland, 1987, p.192) have pointed out that if the best estimate of α is $\hat{\alpha}$, the best estimate of α' should be $m\hat{\alpha} + n$; correspondingly for β'. If 'best' is defined in terms of a least squares estimation method

that is based on vertical distances, then this expectation is satisfied, a fact that advocates rightly regard as a merit of that approach. However, the same argument could be made for a least squares method based on horizontal distances, so it is inconclusive. More seriously, the argument is self-defeating, for unless linear transformations could be shown to be special in some appropriate way, non-linear transformations of y should also lead to correspondingly transformed least squares curves; but this is not in general the case. Indeed, it is practically a universal rule in classical estimation—though maximum likelihood estimation is an exception (see below)—that if $\hat{\alpha}$ is its optimal estimate of α, then $f(\hat{\alpha})$ is suboptimal relative to $f(\alpha)$; in other words, classical estimators are generally not 'invariant'.

Weisberg (1980, p. 214) illustrated this rule with some data on the average brain and body weights of different animal species. The least squares line for the data (expressed logarithmically) was calculated to be $\log(br) = 0.93 + 0.75\log(bo)$. Weisberg then used this equation to estimate (or predict) the log(brain weight) of a so far unobserved species whose mean body weight is 10 kg, obtaining the value 1.68 kg ($0.93 + 0.75\log(10) = 1.68$). Weisberg regarded this as a satisfactory prediction, for one reason, because it is unbiased, and so satisfies a classical criterion for 'good' estimation (see Chapter 5). Now, from the prediction, it seems natural to infer that 47.7 kg (47.7 being the antilogarithm of 1.68) would be a satisfactory prediction for the average brain weight of the species. Yet Weisberg described that conclusion as "naïve", because estimates of brain weights derived in this way are "biased and will be, on the average, too small".

But it seems commonsensical and not at all naive to call for estimates to be invariant under functional transformation; indeed, this view is often endorsed in classical texts (for example by Mood and Graybill, p. 185) when maximum likelihood estimation is under discussion and invariance is offered as one of the properties of that approach which particularly commend it.

So much for intuitive arguments for least squares. They are at best inconclusive, and most classical statisticians believe a more telling case can be made on the basis of certain objective statistical properties of least squares estimates, as we shall describe.

The Gauss-Markov Justification

The objective properties we are referring to are those of unbiasedness, relative efficiency, and linearity, each of which is possessed by the standard least squares estimators of the linear regression parameters. The unbiasedness criterion is well established in all areas of classical estimation, though, as we have already argued, with questionable credentials. The criterion of relative efficiency as a way of comparing different estimators by their variances is also standard; least squares estimates are the most efficient amongst a certain class of estimators, a point we shall return to. Linearity, by contrast, is a novel criterion that was specially introduced into the regression context. It is simply explained: the estimates, $\hat{\alpha}$ and $\hat{\beta}$, of the two linear regression parameters, are said to be *linear* when they are weighted linear combinations of the y_i, that is, when $\hat{\alpha} = \Sigma a_i y_i$ and $\hat{\beta} = \Sigma b_i y_i$. A comparison with the formulas given above confirms that least squares estimates are linear, with coefficients

$$b_i = \frac{x_i - \bar{x}}{c}$$

and

$$a_i = \frac{\Sigma\left(\frac{x_i^2}{n} - \bar{x}x_i\right)}{c}, \text{ where } c = \Sigma(x_i - \bar{x})^2.$$

We now come to the Gauss-Markov theorem, which is the most frequently cited "theoretical justification" (Wonnacott and Wonnacott, p. 17) for the method of least squares. The theorem was first proved by Gauss in 1821, Markov's name becoming attached to it because, in 1912, he published a version of the proof which Neyman believed to be an original theorem (Seal 1967, p. 6). The theorem states that *within the class of linear, unbiased estimators of β and α, least squares estimators have the smallest variance.*

In Chapter 5 we explored the arguments advanced for unbiasedness and relative efficiency as criteria for estimators. Even those who are unconvinced by our objections against them would still have to satisfy themselves that linearity was a reasonable

requirement, before calling on the Gauss-Markov theorem in support of least squares estimation. Often, this need seems not to be perceived, for instance by Weisberg (1980, p. 14), who said, on the basis of the Gauss-Markov theorem, that "if one believes the assumptions [of linear regression], and is interested in using linear unbiased estimates, the least squares estimates are the ones to use"; he did not attend to the question of whether the "interest" in linear estimators which he assumed to exist amongst his readers is reasonable. Seber (1977, p. 49), on the other hand, did describe linear estimators as "reasonable", but gave no supporting argument.

The only justification we have seen for requiring estimators to be linear turns on their supposed "simplicity". "The property of linearity is advantageous", say Daniel and Wood (1980, p. 7), "in its simplicity". Wonnacott and Wonnacott restricted themselves to linear estimates "because they are easy to compute and analyse" (p. 31). But *simplicity and convenience are not epistemological categories; the simplest and easiest road might well be going in the wrong direction and lead you astray.*

Mood and Graybill did not appeal to simplicity but referred mysteriously to "some good reasons why we *would* want to restrict ourselves to linear functions of the observations y_i as estimators" (1963, p. 349); but they compounded the obscurity of their position by adding that "there are many times we would *not*", again without explanation, though they probably had in mind cases where some data points show very large deviations from the presumed regression line or in some other way appear unusual (we shall deal separately with this concern in 7.f below).

The limp and inadequate remarks we have reported seem to constitute the only defences which the linearity criterion has received, and this surely suggests that it lacks any epistemic basis. Mood and Graybill confirmed this impression when they admitted that since "it will not be possible to examine the 'goodness' of the [least squares] estimator $\hat{\beta}$, relative to all functions . . . *we shall have to* limit ourselves to a subset of all functions" (p. 349; italics added). As an example of such a subset they cited linear functions of the y_i and then showed how, according to the Gauss-Markov theorem, the method of least squares excels within that subset, on account of its minimum variance. But

how can we tell whether least squares estimation is not just the best of a bad lot?

The Gauss-Markov argument for using a least squares estimator is that within the set of linear unbiased estimators, it has minimal variance. The argument depends, of course, on minimum variance being a desirable feature. But this presents a further difficulty. For, whatever that minimum variance is in any particular case, there would normally be alternative, biased and/or non-linear estimators of the parameters, whose variance is smaller. And to establish least squares as superior to such alternative methods, you would need to show that the benefit of the smaller variance offered by the alternatives was outweighed by the supposed disadvantage of their bias and/or non-linearity. But this has not been shown, and we would judge the prospects for any such demonstration to be dim.

We conclude that the Gauss-Markov theorem provides no basis for the least-squares method of estimation. Let us move to the third way that the method is standardly defended.

The Maximum Likelihood Justification

The maximum likelihood estimate of a parameter θ, relative to data d, is the value of θ which maximizes the probability $P(d \mid \theta)$. Many classical statisticians regard maximum likelihood estimates as self-evidently worthy, and rarely offer arguments in their defence. Hays is an exception. He argued that the maximum likelihood principle reflects a "general point of view about inference", namely, that "true population situations should be those making our empirical results likely; if a theoretic situation makes our obtained data have very low prior likelihood of occurrence, then doubt is cast on the truth of the theoretical situation" (1969, p. 214; italics removed). And Mood and Graybill pointed to certain "optimum properties" of maximum likelihood estimates, principally that they are invariant under functional transformation, so that, for example, if $\hat{\theta}$ is the maximum likelihood estimate of θ, then $f(\hat{\theta})$ is the maximum likelihood estimate of $f(\theta)$, provided the function has a single-valued inverse. These do not constitute adequate defences. Indeed, there can be no adequate defence, because the maximum likelihood method of estimation cannot be

right in general. This is clear from the fact, to which we alluded before, that any experimental data will be endowed with the maximum probability possible, namely, probability 1, by infinitely many theories, most of which will seem crazy or blatantly false.

In order to apply the maximum likelihood method to the task of estimating α and β in a linear regression equation, the form of the error distribution must be known, for only then can the probability of the data relative to specific forms of the regression be calculated and compared. Least squares, by contrast, can be applied without that knowledge. It turns out that when the regression errors are normally distributed, the maximum likelihood estimates of α and β are precisely those arrived at by least squares, a fact that is often claimed to provide "another important theoretical justification of the least squares method" (Wonnacott and Wonnacott, p. 54).

Those who regard this as a justification clearly assume that the maximum likelihood method is correct, and that it furnishes reasonable estimates; they then argue that since least squares and maximum likelihood coincide in the specific case where the errors are distributed normally, the least squares principle gives reasonable estimates in general. Although we have done so, it is unnecessary to take a view on the merits of maximum likelihood estimation to appreciate that this argument is a non sequitur; you might just as well argue that because you get the right answer with 0 and 1, x^3 is a good way of estimating x^2.

Summary

The classical arguments in favour of least squares estimation of linear regression parameters are untenable. That based on intuition is inconclusive, vague, and lacking in epistemological force. The Gauss-Markov justification rests on the linearity criterion, which itself is unsupported. And the maximum likelihood defence is based on a fallacy.

The two theoretical defences of least squares are standardly cited together as if they were mutually reinforcing or complementary. In fact, the reverse is the case. For while the Gauss-Markov justification presupposes that estimators must be unbiased, the maximum likelihood principle does not. Indeed, it sometimes delivers biased estimates, as for example, in the case at hand: the maximum likelihood estimate of σ^2 is $\frac{1}{n}\Sigma(y_i - \alpha - \hat{\beta}x_i)^2$. But when the regression of y on x is linear, this estimate is biased and needs to be increased by the factor $\frac{n}{n-2}$ to unbias it. (It is this modified estimate, labelled $\hat{\sigma}$ that is always employed in classical expositions.)

The two main 'theoretical' defences of least squares estimation are therefore separately defective and mutually destructive.

But all this does not mean that the least-squares method is entirely wrong. Its great plausibility does have an explanation, a Bayesian explanation. For it can be shown that when you restrict the set of possible regressions to the linear, and assume a normal distribution of errors and a uniform prior distribution of probabilities over parameter values, the least squares and the Bayes solutions coincide, in the sense that the most probable value of the regression parameters and their least squares estimates are identical. When the assumption of a uniform distribution is relaxed, the same result follows, though it is reached asymptotically, as the size of the sample increases (Lindley and El-Sayyad 1968).

7.d Prediction

Prediction Intervals

Fitting a regression curve to data is often said to have as its main practical purpose the prediction of so far unobserved points. We will consider predictions of y values from pre-selected values of x, which is a prediction problem widely considered in statistics texts. Such predictions would be straightforward if one knew the true regression equation, as well as the form of the error distribution. For then the distribution of y_0, the y corresponding to some particular x_0, would also be known, thus enabling one to calculate the probability with which y_0 would fall within any given range. Such a range is sometimes called a *prediction interval.*

The problem is that the regression parameters are usually unknown and have to be estimated from data. To be sure, if those estimates were qualified by probabilities, as they would be after a Bayesian analysis, you could still calculate the probability relative to the data that y_o lay in any specified range. But this option is closed to classical statistics, which proceeds from a denial that theories (parameter values, in the present case) do have probabilities, except in special circumstances that do not prevail in the regression problems we are considering.

Prediction by Confidence Intervals

The classical way round this difficulty is to make predictions using confidence intervals, by a method similar to the one already discussed (in Chapter 5). First, an experiment is imagined in which a fixed set of xs, $x_1, \ldots, x_m$, is chosen. A prearranged number, n ($\geq m$), of x,y readings is then made. A linear regression of y on x is assumed and $\hat{\alpha}$ and $\hat{\beta}$ are the resulting least squares estimates of the corresponding regression parameters. A new x value, x_o, is now considered, with a view to predicting the corresponding y_o. On the linear regression assumption, y_o is a random variable given by $y_o = \alpha + \beta x_o + \in$, with constant variance, σ^2. The following analysis requires the error terms to be normally distributed. A new random variable, u, with variance σ_u^2, is now defined as $u = y_o - \alpha - \beta x_o$. The random variable, t, is then considered, where

$$t = \frac{\sigma u}{\sigma_u - \hat{\sigma}\sqrt{\frac{n-2}{n}}}.$$

The ratio $\frac{\sigma}{\sigma_u}$ is independent of σ, being determined just by $x_1, \ldots, x_m, x_o$, and n, so t can always be computed from the data. Because t has a known distribution (the t-distribution with $n-2$ degrees of freedom), standard tables can be consulted to find specific values, t_1 and t_2, say, which enclose, say, 95 percent of the area under the distribution curve. This means that with probability 0.95, $t_1 \leq t \leq t_2$, from which it follows (Mood and Graybill 1963, pp. 336–37) that

$$P(\hat{\alpha} + \hat{\beta}x_o - \mathbf{A}t_1 \leq y_o \leq \hat{\alpha} + \hat{\beta}x_o + \mathbf{A}t_2) = 0.95,$$

where **A** is a complicated expression involving n, x_o, $\bar{x}$ (the mean of $x_1, \ldots, x_m, x_o$), and σ. The terms $\hat{\alpha}$, $\hat{\beta}$, **A**, and y_o are all random variables that can assume different values depending on the result of the experiment described earlier. If $\hat{\alpha}'$, $\hat{\beta}'$, and **A**′ are the values taken by the corresponding variables in a particular trial, the interval $| \hat{\alpha}' + \hat{\beta}'x_o - \mathbf{A}'t_1, \hat{\alpha}' + \hat{\beta}'x_o + \mathbf{A}'t_2 |$ is a 95 percent confidence interval for y_o. As we noted in our earlier discussion, other confidence intervals, associated with other probabilities, may be described too.

On the classical view, a confidence interval supplies a good and objective prediction of y_o, independent of subjective prior probabilities, with the confidence coefficient (95 percent in this case) measuring how good it is. Support for that view is sometimes seen (for instance by Mood and Graybill 1963, p. 337) in the fact that the width of any confidence interval for a prediction increases with $x_o - \bar{x}$. So according to the interpretation we are considering, if you wished a constant degree of confidence for all predictions, you would have to accept wider, that is, less precise or less accurate intervals, the further you were from $\bar{x}$. This, it is suggested, explains the "intuition" that "we can estimate most accurately near the 'centre' of the observed values of x" (Kendall and Stuart 1979, p. 365). No doubt this intuition is widely shared and is reasonable, provided, of course, that we have presupposed a linear regression, and the explanation given is a point in favour of the confidence interval approach. However, it is insufficient to rescue that approach from the radical criticisms already made. As we explained earlier, the two standard interpretations of confidence intervals—the categorical-assertion interpretation and the subjective-confidence interpretation—are both incorrect. Hence, confidence intervals do not constitute estimates; for the same reasons, they cannot properly function as predictions. That means that the declared principal goal of regression analysis—prediction—cannot be achieved by classical means.

Making a Further Prediction

Suppose, having 'predicted' y_o for a given x_o, its true value is disclosed, thus augmenting the data by an additional point, and

suppose you now wished to make a further prediction of y_0', corresponding to x_0'. It is natural to base the second prediction on the most up-to-date estimates of the linear regression coefficients, which should therefore be recalculated using the earlier data plus the newly acquired data point.

Mood and Graybill in fact recommended such a procedure, but their recommendation did not arise, as it would for a Bayesian, from a concern that predictions should be based on all the relevant evidence. Instead, they said that if the estimated regression equation were not regularly updated in the light of new evidence, and if it were used repeatedly to make confidence-interval predictions, then the basis of the confidence intervals would be undermined. For those confidence intervals are calculated on the assumption that $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}$ are variable quantities, arising from the experiment described earlier (in the previous subsection), and "if the original estimates are used repeatedly (not allowed to vary) the statement may not be accurate" (Mood and Graybill 1963, p. 337).

But Mood and Graybill's idea of simply adding the new datum to the old and re-estimating the regression parameters does not solve the problem. For although it ensures variability for the parameter estimates, it fixes the original data points, which are supposed to be variable, and it varies n, which should be fixed. This means that the true distribution of t, upon which the confidence interval is based, is not the one described above. The question of what the proper t-distribution is has not, so far as we are aware, been addressed. But until it is, classical statisticians ought properly to restrict themselves to single predictions, or else change their methodology, for this inconvenient, unintuitive, and regularly ignored restriction does not arise for the Bayesian, who is at liberty to revise estimates with steadily accumulating data: indeed, there is an obligation to do so.

7.e Examining the Form of a Regression

The classical way of investigating relationships between variables is, as we have described, to assume some general form for the relationship and then allow the data to dictate its particular shape by supplying values for the unspecified parameters. In their starting assumption, statisticians show a marked preference for linear regressions, partly on account of the conceptual simplicity of such regressions and partly because their properties with respect to classical estimation techniques have been so thoroughly explored. But statistical relationships are not necessarily nor even normally linear ("we might expect linear relationships to be the exception rather than the rule"—Weisberg 1980, p. 126); and any data set can be fitted equally well (however this success is measured) to infinitely many different curves; hence, as Cook (1986, p. 393) has said, "some reassurance [that the model used is "sensible"] is always necessary".

But how is that reassurance secured, and what is its character? The latter question is barely addressed by classical statisticians, but the former, it is widely agreed, can be dealt with in one or more of three ways. First, certain aspects of possible models, it is said, can be checked using significance tests. There is, for example, a commonly employed test of the hypothesis that the β-parameter of a linear regression is zero. This employs a t-test-statistic with $n-2$ degrees of freedom: $t = \dfrac{\hat{\beta}}{s(\hat{\beta})}$, where $s(\hat{\beta})$ is the standard error of $\hat{\beta}$, which can be estimated from the data. Another test is often employed to check the homoscedasticity assumption of the linear regression hypothesis, that is, the assumption that σ is independent of x. Such tests are of course subject to the strictures we made on significance tests in general, and consequently, in our view, they have no standing as modes of inductive reasoning. But even if the tests were valid, as their advocates maintain, they would need to be supplemented by other methods for evaluating regression models, for the familiar reason that the conclusions that may be drawn from significance tests are too weak for practical purposes. Learning that some precise hypothesis, say that $\beta = 0$, has been rejected at such-and-such a significance level is to learn very little, since in most cases it would already have been extremely unlikely, intuitively speaking, that the parameter was *exactly* zero. In any case, most practitioners require more than this meagre, negative information; they wish to know what

the true model actually is. The methods to be discussed in the next two subsections are often held to be helpful in this respect. They deal first with the idea that the preferred regression model or models should depend on prior knowledge; and secondly with techniques that subject the data to detailed appraisal, with a view to extracting information about the true model.

Prior Knowledge

Hays (1969, p. 551) observed that "when an experimenter wants to look into the question of trend, or form of relationship, he has some prior ideas about what the population relation should be like. These hunches about trend often come directly from theory, or they may come from the extrapolation of established findings into new areas."

Such hunches or prior ideas about the true relationship are often very persuasive. For example, in studying how the breaking load *(b)* varies with the diameter *(d)* of certain fibre segments, Cox (1968, p. 268) affirmed that as the diameter vanishes, so should the breaking load; Seber (1977, p. 178) regarded this as "obvious". More often, prior beliefs are less strongly held. For example, Wonnacott and Wonnacott, having submitted their data on wheat yields at different concentrations of fertiliser to a linear least-squares analysis, pointed out that the assumption of linearity is probably false: *"it is likely* that the true relation increases initially, but then bends down eventually as a 'burning point' is approached, and the crop is overdosed" (p. 49; italics added). Similarly, Lewis-Beck observed that although a linear relation between income and number of years of education appears satisfactory, judged from the data, *"it seems likely* that relevant variables have been excluded, for factors besides education undoubtedly influence income" (1980, p. 27; italics added). Other statisticians express their uncertainty about the general model in less obviously probabilistic terms, such as "sensible" and "reasonable on theoretical grounds" (Weisberg 1980, p. 126), and "plausible" (Atkinson 1985, p. 3); while Brook and Arnold talked of "theoretical clues . . . which *point to* a particular relationship" (1985, p. 12; italics added); and Sprent (1969, p. 120), stretching tentativeness almost to the limit, said that an "experimenter is often in a position . . . to decide that it is reasonable to assume that certain general types of hypothesis . . . may hold, although he is uncertain precisely which".

One further example: Cox (1968, p. 268) argued that if, in the case of the breaking load and the diameter of fibres, the two lines $b \propto d$ and $\log b \propto \log d$ fit the data equally well, then the "second would in general be preferable because . . . it permits easier comparison with the theoretical model load $\propto$ (diameter)^2". This is a very circumspect way of recommending a regression model and would seem to be no recommendation at all unless the "theoretical model" were regarded as likely to be at least approximately true. That this is the implicit assumption seems to be confirmed by Seber's exposition (1977, p. 178) of Cox's view, in which he commended the model slightly less tentatively, saying that it "might be" a "reasonable assumption".

One might expect the natural uncertainty attaching to the general model to be revised, ideally diminished, by the data, so producing an overall conclusion that incorporates both the prior and posterior information. This is how a Bayesian analysis would operate. A Bayesian would interpret the various expressions of uncertainty uniformly as prior probabilities and then use the data to obtain corresponding posterior probabilities, though if the distribution of prior beliefs is at all complicated, the mathematics may become difficult or even intractable.

In the classical case, the difficulty is not merely mathematical or technical, but arises from a fundamental flaw. For, in the first place, the prior evaluation of a model in the light of plausible theories and previous data seems not, as a rule, to be objectively quantifiable, as the classical approach would demand; certainly none of the authors we have quoted offers any measure, objective or otherwise, of the strength of their hunches. Secondly, even if the reasonableness of a theory could be objectively measured, classical statistics offers no way of combining such measures with the results of standard inference techniques (significance testing, confidence-interval estimation, and so forth) to achieve an aggregate index of appraisal.

The difficulty of incorporating uncertainty about the model into standard classical inferences is apparent from a commonly encountered discussion concerning predictions. Typical exposi-

tions proceed by first assuming a linear relation. The apparatus of confidence intervals is next invoked as a way of qualifying predictions with a degree of confidence. It is then explained that the linearity assumption is often doubtful, though perhaps roughly correct over the relatively narrow experimental range, and hence that one should not expect predictions from an equation fitted by least squares to be quite accurate (for example Weisberg 1980, p. 126; Seber 1977, p. 6; Gunst and Mason 1980, pp. 56–63; Wonnacott and Wonnacott 1980, p. 49).

But instead of measuring this uncertainty about the model and then amalgamating it with the uncertainty reflected in the confidence interval, so as to give an overall level of confidence in a prediction, classical statisticians merely issue warnings to be "careful not to attempt to predict very far outside the [range covered by the data]" (Gunst and Mason, p. 62), and to "be reluctant to make predictions [except for] . . . new cases with predictor variables not too different from [those] . . . in the construction sample" (Weisberg, p. 215). But these vague admonitions do not signify how such caution should be exercised (should one hold off from predicting altogether, or tremble slightly when hazarding a forecast, or what?).

There is also the question of how close x_0 should be to the x values in the data, before the corresponding y_0 can be predicted with assurance. This is always given an arbitrary answer. For example, Weisberg, in dealing with the case where y is linearly related to two predictor variables x and z, suggested, with no theoretical or epistemological sanction, that a "range of validity for prediction" can be determined by plotting the data values of x and z and drawing "the smallest closed figure that includes all the points" (p. 216). Making such a drawing turns out to be difficult, but Weisberg considered that "the smallest volume ellipsoid containing these points" (p. 216) would be a satisfactory approximation to the closed figure. However, he is uneasy with the proposal since, in the example on which he was working, there was "a substantial area inside the [ellipsoid] . . . with no observed data" (p. 217) and where, presumably, he felt prediction is unsafe. Weisberg's demarcation between regions of safe and risky predictions and his suggested approximation to it have some plausibility, but they seem quite arbitrary from the classical point of view.

Classical statisticians are caught in a cleft stick. If they take account of plausible prior beliefs concerning the regression model, they cannot properly combine those beliefs with the classical techniques of inference. On the other hand, if they use those techniques but eschew prior beliefs, they have no means of selecting, arbitrary stipulation apart, among the infinitely many regression models that are compatible with the data.

Data Analysis

A possible way out of this dilemma is to abandon the imprecise, unsystematic, and subjective appraisals described in the previous section and rely solely on the seemingly more objective methods of 'data analysis', though most statisticians would not go so far, preferring to avail themselves of both techniques. Data analysis, or 'case analysis', as it is often called, is an attempt to discriminate in a non-Bayesian way between possible regression models, through a close examination of the individual data points. There are three distinct approaches to data analysis, which we shall consider in turn.

i Inspecting scatter plots. The simplest kind of data analysis involves visually examining ordinary plots of the data ('scatter plots') in an informal way. The procedure was authoritatively endorsed by Cox (1968, p. 268)—"the choice [of relation] will depend on preliminary plotting and inspection of the data"; and it was "strongly recommended" by Kendall and Stuart (1979, p. 292), because "it conveys quickly and simply an idea of the adequacy of the fitted regression lines". Visual inspection is widely employed in practice to augment the formal process of classical estimation. Weisberg (p. 3), for instance, motivated a linear regression analysis of certain data on the boiling points of water at different atmospheric pressures by referring to the "overall impression of the scatter plot . . . that the points generally, but not exactly, fall on a straight line". And Lewis-Beck (p. 15) claimed that "visual inspection of [a certain] . . . scatter plot suggests the relationship is essentially linear".

Weisberg (p. 99) argued that the data of Diagram 1 showed a pattern that "one might expect to observe if the simple linear

regression model were appropriate". On the other hand, Diagram 2 "suggests . . . [to him] that the analysis based on simple linear regression is incorrect, and that a smooth curve, perhaps a quadratic polynomial, could be fit to the data . . . ". Although the data are different in the two cases, they give the same least squares lines, as the diagrams below illustrate.

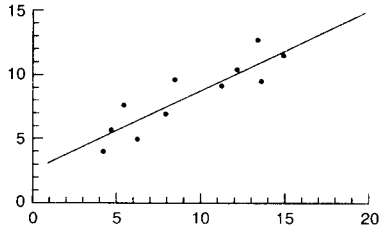


DIAGRAM !

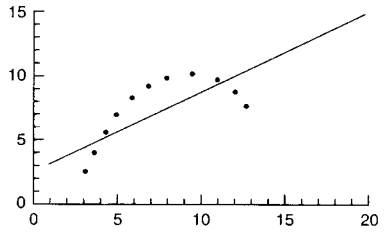


DIAGRAM 2

Probably most scientists would agree with the various judgments that statisticians make on the basis of visual inspections of scatter diagrams. But how are such judgments arrived at? It could be, indeed, it seems likely, that a close analysis would reveal a Bayesian mechanism, but it is hard to imagine how any of the standard classical techniques could be involved in either explaining or justifying the process, nor do classical statisticians claim they are. The whole process is impressionistic, arbitrary, and subjective.

ii Outliers. Another instructive data pattern is illustrated in Diagram 3. One of the points stands much further apart from the least squares line than any other and this suggests to Weisberg (p. 99) that "simple linear regression may be correct for most of the data, but one of the cases is too far away from the fitted regression line". Such points are called *outliers*.

Outliers are defined, rather imprecisely, as "data points [with] . . . residuals that are large relative to the residuals for the remainder of the observations" (Chatterjee and Price 1977, p. 19). (The residual of a data point is its vertical distance from a fitted line or curve.) As is often noted, a point may be an outlier for three distinct reasons. First, it could be erroneous, in the sense that it

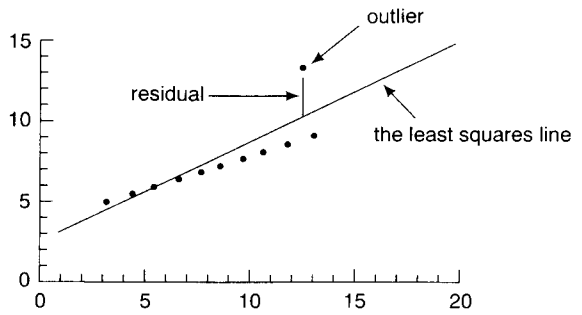


DIAGRAM 3

resulted from a recording or transcription error or from an improperly conducted experiment; an outlier could also arise because the assumed regression model is incorrect; on the other hand, the model might be correct and the outlier be simply one of those relatively improbable cases that is almost bound to occur sometimes. Suppose that careful checking has more or less excluded the first possibility, does the outlier throw any light on whether an assumed regression equation is correct or not? This is a question to which classical statisticians have applied a variety of non-Bayesian methods, though, as we shall argue, without any satisfactory result.

Some authors, for example Chatterjee and Hadi, seem not to take seriously the possibility that the linear least-squares line is wrong, when they note that an outlier in their data, if removed, would hardly affect the fitted line and conclude from this that "there is little point in agonizing over how deviant it appears" (1986, p. 381). But this conclusion is unjustified, and was not endorsed by Chatterjee when he previously collaborated with Price.

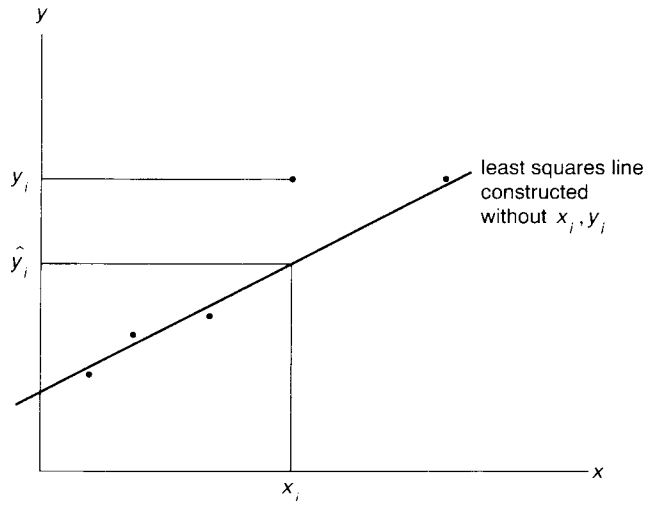
Chatterjee and Price's (1977) approach may be illustrated with their own example. Their data consisted of 30 x,y readings, the nature of which need not concern us. They applied the linear least squares technique to the readings, obtaining an upwardly sloping regression line. They next examined a number of data plots, first y against x, then residuals against x, and finally the residuals against $\hat{y}$ ($\hat{y}_i$ is the point on the fitted line corresponding to x_i). Four of the points in their data stood out as having particularly

large residuals; moreover, a visual inspection of the various plots made it "clear [to the authors] that the straight line assumption is not confirmed" (p. 24), though it "looks acceptable" in the middle of the x range. On this informal basis, Chatterjee and Price concluded tentatively that the true line has a zero gradient and that y is independent of x. They checked this conjecture by dropping the four outliers, computing a new least-squares line and then examining the revised plots of residuals. This time they found no discernible pattern; from a casual inspection "the residuals appear to be randomly distributed around [the line] e[residual] = 0" (p. 25). Unfortunately, the conclusion from all this is disappointingly weak. It is that the regression with zero gradient "is a satisfactory model for analyzing the . . . data, *after the deletion of the four points"* (p. 25; italics added). But this says nothing about the true relation between x and y. Indeed, the authors acknowledged that this question was still open by then asking: "what of the four data points that were deleted?" They repeated the truism that the points may be "measurement or transcription errors", or else "may provide more valuable information about the . . . relationships between y and x". We take the latter to mean that the line derived from the data minus the outliers might be wrong, perhaps, it should be added, wildly so (of course, it might be right, too). Whether the fitted line is right or wrong, and if wrong, what the true line is, are questions that Chatterjee and Price simply left to further research: "it may be most valuable to try to understand the special circumstances that generated the extreme responses" (p. 25). In other words, their examination of the data from different points of view has been quite uninformative about the true regression and about why some of the points have particularly large residuals relative to the linear least-squares line.

Chatterjee and Price called theirs an "empirical approach" to outliers, since it takes account not just of the sizes of the residuals but also of their patterns of distribution. They disagreed with those who use significance tests alone to form judgments from outliers since, in their view, "[i]t is a combination of the actual magnitude and the pattern of residuals that suggests problems" (p. 27). But although significance tests take little account of the pattern of the results, compared with the empirical approach, they are more in keeping with the objectivist ideals of classical statistics

and offer more definite conclusions about the validity of hypothesized regression equations.

Weisberg is a leading exponent of the application of significance tests in this context. His method is to perform a linear least-squares analysis on the data set minus one point (in practice this would be an outlying point) and then to use a significance test to check the resulting regression equation against the removed point. To see how this 'outlier test' is carried out, suppose the fitted line derived from the data, minus the ith point, has parameters $\hat{\alpha}_{-i}$ and $\hat{\beta}_{-i}$. The assumption that this is the true line will be the null hypothesis in what follows. Let $\hat{y}_i$ be the y value corresponding to x_i on the null hypothesis line.



Provided the data were properly collected and recorded, it is natural to expect that the larger the discrepancy between $\hat{y}_i$ and y_i, the less would be one's confidence that the line drawn is correct. This idea is given classical clothes through a corresponding significance test. Weisberg's test uses as test-statistic a particular function of $y_i - \hat{y}_i$ and other aspects of the data—this has the t-distribution with $n - 3$ degrees of freedom. If the t-value recorded for a data set, relative to a particular point (x_i, y_i) in that set, exceeds the test's critical value, then the null hypothesis would be rejected at the corresponding level of significance.

But fixing the appropriate critical value is interestingly problematic and shows up the pseudo-objectivity and ineffectiveness of the whole procedure. It is required in significance testing to ensure that the null hypothesis would be falsely rejected with some pre-designated probability, usually 0.05 or 0.01. Now Weisberg's test could be conducted in a spectrum of ways. At one end of the spectrum you could decide in advance to perform just one significance test, using the single datum corresponding to a pre-specified x_i (call this the first testing plan). At the other extreme, the plan could be to check the significance of every data point. As Weisberg points out, if the test were, as a matter of policy, restricted to the datum with the largest t-value, one would in effect be following the second testing plan.

Suppose you selected a significance level of 0.05. This would be classically acceptable, provided the first plan was adopted and the test applied just to a single, pre-selected point. But if the second plan was adopted, and the same significance level chosen, the *overall* probability of rejecting the null hypothesis, if it were true, would be a multiple of 0.05, that multiple depending on the number of points in the data set. You would, in this case, then have to reduce the significance level of the individual tests, in order to bring the overall significance level to a classically acceptable level. (In Weisberg's example, involving 65 data points, the second testing plan would need to use a significance level of 0.00077 for the individual tests.)

So whether a particular t-value is significant and warrants the rejection of a null hypothesis is sensitive to how the person performing the significance test planned to select data for testing. But, as we have stressed before, such private plans have no epistemic significance, and without justifiable, public rules to fix the most appropriate plan, the present approach is subject to personal idiosyncrasies that are at odds with its supposed objectivity.

Weisberg (p. 116) does in fact propose a rule for choosing a testing plan, namely, that the one involving a single significance test, should be adopted only "if the investigator suspects in advance" that a particular case will fail to fit the same regression line as the others. But what should the experimenter do if, having decided to adopt this plan, the anticipated outlier unexpectedly fitted the pattern and some of the other points were surprisingly

prominent outliers? And why should informal and untested "suspicions" have any standing in this area?

Weisberg introduced and illustrated his rule with his data on the average brain (*br*) and body weights (*bo*) of different species of mammals, which show man as the most prominent outlier in a plot of log(*br*) against log(*bo*). Weisberg (p. 130) argued that since "interest in Man as a special case is a priori", the outlier test should employ the larger significance level, corresponding to the first testing plan. On this basis the datum on man was significant and so Weisberg concluded that "humans would be declared to have brain weight that is too large to be consistent [*sic*] with the [log-linear] model for the data". But if there had been no prior "interest in Man", the datum would not have been significant and Weisberg would have reached the opposite conclusion.

Weisberg's example, moreover, does not even conform to his own rule: man is not picked out because of an earlier "suspicion" that he is an exception to the log-linear rule, but because of a prior "interest" in Man as a special case. No explanation is given for this change, though we conjecture that it is made because, first, from a visual inspection of the data, Man seems clearly to be an exception to a log-linear rule; secondly, if the smaller significance level, corresponding to the second testing plan, were employed in Weisberg's outlier test, Man would *not* be significant and the linear rule would, counter-intuitively, have to be accepted; and thirdly, there is no plausible reason to suspect Man a priori of being an exception to a log-linear rule—hence, no reason to employ the larger critical value in the significance test. This may perhaps explain why Weisberg changed his rule to suit these circumstances, but clearly it provides no justification.

It seems undeniable that the distribution patterns of data, including the presence of outliers, often tell us a great deal about the regression, but they seem not to speak in any known classical language. It is perhaps too early to say for certain that the medium of communication is Bayesian, since a thorough Bayesian analysis of the outlier notion is still awaited. The form that that analysis will take is perfectly clear, however. As with every other problem concerning the evaluation of a theory in the light of evidence, it will involve the application of Bayes's theorem. Thus, for a Bayesian, regression analysis is not divided into separate

compartments operating distinct techniques, each requiring its own justification; there is just one principle of inference, leading to one kind of conclusion.

iii Influential points. Another data pattern that statisticians often find instructive is illustrated in Diagram 4 (Weisberg, p. 99).

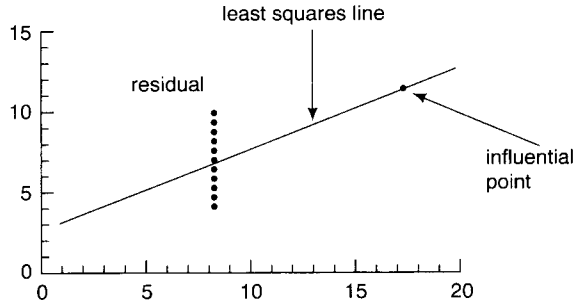


DIAGRAM 4

The isolated point on the right has a nil residual, so the considerations of the previous section give no reason to doubt the linear least squares line. But the point is peculiar in that the estimated line depends very largely on it alone, much more so than on any of the other data. Such points are called 'influential'. Diagram 4 shows an extreme case of an influential data point; it is extreme because without it no least squares solution even exists. There are less pronounced kinds of influence which statisticians also regard as important, where particular data points or a small subset of the data "have a disproportionate influence on the estimated [regression] parameters" (Belsley *et al.* 1980, p. 6).

Many statisticians regard such influential points as deserving special attention, though they rarely explain why; indeed, large textbooks are written on how to measure "influence", without the purpose of the project being adequately examined. All too often, the argument proceeds thus: "these [parameter estimates] . . . can be substantially influenced by one observation or a few observations; that is, not all the observations have an equal importance in least squares regression . . . It is, *therefore,* important for an analyst to be able to identify such observations and assess their effects on various aspects of the analysis" (Chatterjee and Hadi

1986, p. 379; italics added). Atkinson (1986, p. 398) and Cook (1986, p. 393) give essentially the same argument, which is clearly a non sequitur, without the further and no doubt implicit assumption that conclusions obtained using influential points are correspondingly insecure. Weisberg (p. 100) is one of the few who states this explicitly, when he says that "[w]e must distrust an aggregate analysis that is so heavily dependent upon a single case". Belsley *et al.* (1980, p. 9) say roughly the same but express it in a purely descriptive mode: "the researcher is likely to be highly suspicious of the estimate [of the slope of a regression line]" that has been obtained using influential data.

But why should we distrust conclusions substantially based on just a few data points? Intuitively, there are two reasons. The first is this: the true regression curve passes through the mean, or expected value, of y at each x. In describing our best guess as to that curve, and hence our best estimate of those means, we are guided by the observed points. But we are aware that any of the points could be quite distant from the true regression line, either because the experiment was badly conducted or because of an error in recording or transcribing, or simply because it is one of those rare and improbable departures from the mean that is almost bound to occur from time to time. Sharp departures of a point from its corresponding mean are relatively improbable, but if such a discrepancy actually arose and the datum was, moreover, influential, the conclusion arrived at would be in error by a corresponding margin. Our intuition is that the more likely a reading is to be discrepant in the indicated sense, and the more influential it is, the less certain would one be about the regression. The second intuitive reason for distrusting conclusions that are partly derived from influential data applies when they are separated from the rest of the data by a wide gap, as for example in Diagram 4; intuitively we feel that in the range where there are few or no data points, we cannot be at all sure of the shape of the regression relation.

A programme of research that was initiated in the early 1970s aims to explicate these intuitions in classical (or at any rate, non-Bayesian) terms and to develop rules to guide and justify them. This now flourishing field is called 'influence methodology', its chief object being to find ways of measuring influence.

The idea governing the measurement of influence is this. You should first estimate a regression parameter using all the data, and then re-estimate it with a selected data point deleted, noting the difference between the two estimates. This difference is then plugged into an 'influence function' or, as it is sometimes called, a 'regression diagnostic', to produce an index of how influential the deleted point was. To put this programme into effect, you need first to decide on the parameters whose estimates are to be employed in the influence measure. If the regression has been assumed linear with a single independent variable, you could choose from among the three parameters, α, β, and σ, as well as from the infinity of possible functional combinations of these. Secondly, a measure of influence is required, that is, some function of the difference noted above and (possibly) other aspects of the data; the choice here is similarly vast. Finally, there has to be some way of using particular values of an influence measure to judge the reliability of the data point or of the regression assumptions.

Many influence functions have been proposed and suggestions advanced as to the numerical value such functions should take before the reading concerned is designated 'influential'. But these proposals and suggestions seem highly arbitrary, an impression confirmed by those working in the field. For instance, the author of a function known as Cook's Distance admitted that "[f]or the most part the development of influence methodology for linear regression is based on ad hoc reasoning and this partially accounts for the diversity of recommendations" (Cook 1986, p. 396).

Arbitrariness and adhocness would not be a feature of this area if a relation had been established between a point's influence as defined by a given measure and the reliability or credibility of a fitted line or curve. No such relation has been established. Nevertheless, recommendations on how to interpret levels of influence are sometimes made. For example, Velleman and Welsch (1981) felt that "values [of their preferred function) greater than 1 or 2 seem reasonable to nominate . . . for special attention"; but they failed to say what special attention is called for, nor why.

Many of those most active in this area acknowledge the apparent absence of any objective epistemic constraints when it comes

to interpreting influence measures. For example, Welsch (1986, p. 405), joint author of another influence function, conceded that "[e]ven with a vast arsenal of diagnostics, it is very hard to write down rules that can be used to guide a data analysis. So much is really subjective and subtle". And Velleman (1986, p. 413), who has already been quoted, talked of "the need to combine human judgment with diagnostics", without indicating how this judgment operates, is justified, or coheres with the rest of classical statistics.

A full Bayesian analysis of influential data has yet to be undertaken. It would, we are sure, endorse many of the intuitions that guide non-Bayesian treatments, but unlike these, the path it should take and its epistemic goal are clear; using Bayes's theorem, it would have to trace the effects of uncertainty in the accuracy of readings and of relatively isolated data points on the posterior probabilities of possible regression models. The technical difficulties facing such a programme are formidable, to be sure (and these difficulties are often referred to by critics wishing to cast doubt over the whole Bayesian enterprise), but the difficulties that any Bayesian analysis of regression data faces arise only from the complexity of the situation and by no means reflect on the adequacy of the methodology. The three-body problem in physics provides an instructive analogy; this problem has so far resisted a complete solution in Newtonian terms and may, for all we know, be intrinsically insoluble, but nobody thinks that Newton's laws are in the least discredited because of the mathematical difficulties of applying them in this area.

By contrast with a Bayesian approach, the programme of influence methodology based on classical ideas runs into trouble not simply because of intractable mathematics but, we suggest, because it has no epistemically relevant goal. This explains why the rules and techniques proposed are arbitrary, unjustified, and ad hoc; it is hard to see how they could be otherwise.

7.f Conclusion

In earlier chapters, we catalogued various aspects of classical methods of testing and estimation that show that they are unsuited to the tasks of inductive inference for which they were invented.

Those same shortcomings, not surprisingly, surface too when regression problems are at issue. However, some new difficulties are also revealed. For instance, extra criteria for estimation are required, which have led to the plausible, but classically indefensible, least squares principle. There are also extra sources of subjectivity, for instance, in selecting the regression model, both when taking account of 'prior knowledge' and when judging models by informally inspecting scatter plots. They are also present in the process of checking models against outliers and influential data. This subjectivity frequently passes unnoticed. Thus Wonnacott and Wonnacott (1980, p. 15) set themselves firmly against judging the gradient and intercept of a regression line "by eye", on account of its subjectivity—"we need to find a method that is objective"; but without batting an eyelid at their inconstancy, they allow personal judgment and the visual inspection of scatter plots to play a crucial part in determining the overall conclusion.

Unlike the hotchpotch of ad hoc and unjustifiable rules of inference that constitute the classical approach, Bayes's theorem supplies a single, universally applicable, well-founded inductive rule which answers what Brandt (1986, p. 407) calls the "most important . . . need for integration of this [influence methodology] and many other aspects of [classical] regression and model fitting into a coherent whole".

CHAPTER 8
Bayesian Induction: Statistical Theories

Bayesian induction is the computation via Bayes's theorem of a posterior probability, or density distribution from a corresponding prior distribution, on receiving new information. The theorem does not discriminate between deterministic and statistical theories, and so affords a uniform treatment for both, in contrast to the hotchpotch of non-Bayesian methods that have been invented to suit different circumstances. In an earlier chapter we considered how non-Bayesian and Bayesian approaches compared when they were applied to deterministic hypotheses. Here we make a similar comparison in relation to statistical theories.

8.a | The Question of Subjectivity

The prior distribution from which a Bayesian analysis proceeds reflects a person's beliefs before the experimental results are known. Those beliefs are subjective, in the sense that they are shaped in part by elusive, idiosyncratic influences, so they are likely to vary from person to person. The subjectivity of the premises might suggest that the conclusion of a Bayesian induction is similarly idiosyncratic, subjective and variable, which would conflict with a striking feature of science, namely, its substantially objective character.

A number of attempts have been made to reconcile Bayesian methodology with scientific objectivity. For example, it has been argued that the difficulty may be stemmed at source by repudiating subjective in favour of purely objective prior probabilities on which all scientists might rationally agree. But the fact is that scientists often take very different views on how credible particular theories are, especially in the early stages of an investigation. And,

more seriously, although the idea that scientific theories possess unique, objectively correct, inductive probabilities is eminently appealing, it has so far resisted determined efforts at a satisfactory analysis, except in the special cases of tautologies and contradictions, and there is now a wide consensus that no such analysis is possible.

A second approach tries to exploit certain limit theorems of Bayesian inductive theory. These are theorems (treated in detail in the next chapter) to the effect that under certain, mild conditions, for instance, that there is agreement on which hypotheses, if any, have zero prior probability, different Bayesian agents will, in the limit, as data accumulate without bounds, agree in their posterior views, whatever their subjective prior opinions were. But, remarkable as they are, the limit theorems are of little utility in the present context. For one thing, since they deal only with the properties of the posterior probability function in the limit, as the amount of data increases to infinity, they say nothing at all about its character after any actual and hence finite amount of data; they are therefore incapable of acting in either a normative or an explanatory role.

There is a third approach, which is the one we favour. It does not seek to account for a global identity of opinion, nor for a convergence to such an identity once some specified amount of data is available, neither of which seem to be universal facts of science. The approach recognizes that scientific opinions often do converge, often indeed, very quickly, after relatively little data. And it appeals to Bayes's theorem to provide an explanatory framework for this phenomenon, and to give specific explanations for specific circumstances. We believe that such explanations should be judged case by case, and that when this is done, the judgments will usually be favourable.

Two cases which crop up frequently in the literature of statistical inference, as well as in practical research, and which we considered earlier in connexion with the frequentist approach, are the estimation firstly, of the mean of a normal population, and secondly, of a binomial proportion. We will show that in both these estimation tasks, Bayesian reasoners, even when their prior opinions are very different, are forced into ever-closer agreement in their posterior beliefs as the data accumulate. And, more significantly, we will see that this convergence of opinion is fairly rapid.

Estimating the Mean of a Normal Population

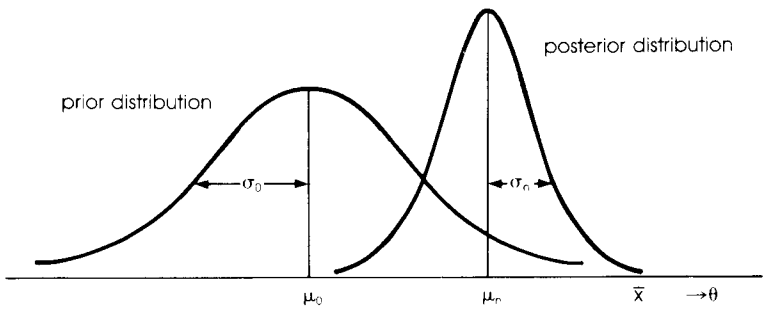
We used this customary example in Chapter 5 to present the ideas of classical interval estimation. In the example, the population whose mean is to be estimated is already known to be normal and known to have standard deviation σ. The assumption of such knowledge is more or less realistic in many cases, for instance, where an instrument is used to measure some physical quantity. The instrument would, as a rule, deliver a spread of results if used repeatedly under similar conditions, and experience shows that this variability, or error distribution, often approximates a normal curve. Making measurements with such an instrument would then be practically equivalent to drawing a random sample of observations from a normal population of possible observations whose mean is the unknown quantity and whose standard deviation was established from previous calibrations.

Let θ be a variable that ranges over possible values of the population mean. We shall assume for the present that prior opinion is represented by a density distribution over θ that is also normal, with a mean of μ_0 and a standard deviation of σ_0. In virtually no real case would the prior distribution be strictly normal, for physical considerations usually impose limits on a parameter's possible values, while normal distributions assign positive probabilities to every range. So for instance, the average height of a human population could neither be negative, nor above five thousand miles. Nevertheless, a normal distribution often provides a mathematically convenient idealization of sufficient accuracy. Assuming a normal distribution simplifies this illustration of Bayesian induction at work, but as we show later, the assumption may be considerably relaxed without substantially affecting the conclusions.

Suppose a random sample of size n and mean $\bar{x}$ has been drawn from the population. The posterior distribution of θ, relative to these data, turns out to be normal, like the prior, and its mean, μ_n, and its standard deviation, σ_n, are given by:

$$\mu_n = \frac{n\bar{x}\sigma^{-2} + \mu_0\sigma_0^{-2}}{n\sigma^{-2} + \sigma_0^{-2}} \quad \text{and} \quad \frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}.$$

These results, which are proved by Lindley, 1965, for example, are illustrated below.



The *precision* of a distribution is defined as the reciprocal of its variance. So the second of the above equations tells us that the precision of the posterior distribution increases with the precision, σ^2, of the population whose mean is being estimated. By the same token, the precision of a measured value increases with that of the measuring instrument (whose precision is the reciprocal of the variance of its error distribution). The more precise an estimate the less uncertainty there is about the parameter's value, and the natural wish to diminish uncertainty accounts for the appeal of the efficiency criterion that classical statisticians have imposed, for inadequate reasons (see 5.f above), on their estimators.

The above equations also show that as n increases, μ_n, the mean of the posterior distribution, tends to $\bar{x}$, the mean of the sample. Similarly, σ_n^2 tends to $\frac{\sigma^2}{n}$, a quantity that depends on the sample and on the population but not on the prior distribution. This means that as the sample is enlarged, the contribution of the prior distribution, and so of the subjective part of the inference, lessens, eventually dwindling to insignificance. Hence two people proceeding from different normal prior distributions would, with sufficient data, converge on posterior distributions that were arbitrarily close. Moreover—and this is the crucial point for explaining the objectivity of the inference—the objective information contained in the sample becomes the dominant factor relatively quickly.

We can show how quick the convergence of opinion may be by an example where the aim is to estimate the mean of a normal population whose standard deviation is already known to be 10.

And we consider the case in which one person's normal prior distribution over the population mean is centred on 10, with a standard deviation of 10, while another's is centred on 100, with a standard deviation of 20. This difference represents a very sharp divergence of initial opinion, because the region that in the view of the first person almost certainly contains the true mean is practically certain not to contain it as far as the second is concerned. But even so profound a disagreement as this is resolved after relatively few observations. The following table shows the means and standard deviations of the posterior distributions of the two people, relative to random samples of various sizes, each sample having a mean of 50.

Sample size	*Person 1*		*Person 2*	
n	μ_n	σ_n	μ_n	σ_n
0	10	10	100	20
1	30	7.1	60	8.9
5	43	4.1	52	4.4
10	46	3.0	51	3.1
20	48	2.2	51	2.2
100	50	1.0	50	1.0

We deliberately chose an extreme example, where the prior distributions scarcely overlap. The first line of the table, corresponding to no data, represents this initial position. Yet we see that a sample of only 20 brings the two posterior distributions very close, while a sample of 100 renders them indistinguishable. Not surprisingly, the closer opinions are at the start, the less evidence is needed to bring the corresponding posterior opinions within given bounds of similarity.[1] Hence, although the Bayesian analysis of the case under consideration must proceed from a largely subjective prior distribution, the most powerful influence on its conclusion is the objective experimental evidence. We shall see that the same is more generally true.

[1] See, for instance, Pratt *et al.* 1965, Chapter 11.

Estimating a Binomial Proportion

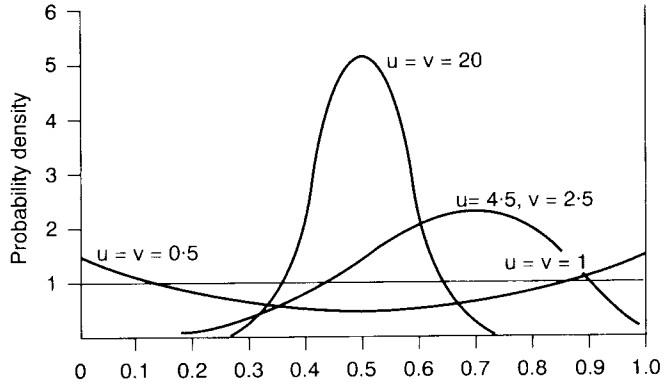
Another standard problem in inferential statistics is how to estimate a proportion, for instance, of red counters in an urn, or of Republican sympathisers in the population, or of the physical probability of a coin turning up heads when it is flipped. Data are collected by randomly sampling the urn or the population, with replacement, or by flipping the coin, and then noting for each counter sampled whether it is red or not, and for each person sampled his or her political sympathies, and for each toss of the coin whether it landed heads or tails. If, as in these cases, there are just two possible outcomes, with probabilities θ, and $1 - \theta$, constant from trial to trial, then the data-generating process is known as a Bernoulli process and θ and $1 - \theta$ are called the Bernoulli parameters, or the binomial proportions. The outcomes of a Bernoulli process are conventionally labelled 'success' and 'failure'.

The Bayesian method of estimating a Bernoulli parameter from given experimental results starts, of course, by describing a prior distribution, which we shall assume to have the form of a so-called *beta* distribution. This restriction has the expository advantage of simplifying the calculation of the corresponding posterior distribution. The restriction is not in fact a severe one, for beta distributions take on a wide variety of shapes, depending on the values of two positive-valued parameters, u and v, enabling you to choose a beta distribution that best approximates your actual distribution of beliefs. And as we shall see, with sufficient data, posterior probabilities are not much affected by even quite big changes in the corresponding priors, so that inaccuracies in the specification of the prior will then not have a significant effect.

A random variable x is said to have the beta distribution if its density is given by

$$\begin{aligned} P(x) &= B(u, v)\, x^{u-1}(1-x)^{v-1} && 0 < x < 1 \\ &= 0 && \textit{elsewhere.} \end{aligned}$$

The parameters u and v are both greater than 0, and $B(u,v) = \dfrac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)}$. We do not need to spell out the gamma function in detail here, except to note that when w is a positive integer that $\Gamma = (w-1)!$ (with 0! defined as 1). When w is non-integral, the value of the gamma function can be obtained from tables in mathematical handbooks. The following diagram illustrates some beta distributions for various values of u and v.



SOME BETA DISTRIBUTIONS

The mean and variance of a beta distribution are given by:

$$\text{mean} = \frac{u}{u+v} \qquad \text{variance} = \frac{\left(\dfrac{u}{u+v}\right)\left(\dfrac{v}{u+v}\right)}{u+v+1}.$$

If the prior distribution over the Bernoulli parameters is of the beta form, Bayes's theorem is particularly easy to apply. For suppose a random sample of n observations derived from a Bernoulli process shows s successes and f failures, it then turns out that the posterior distribution is also of the beta form, with parameters $u' = u + s$ and $v' = v + f$. Hence, the mean of the posterior distribution is $\dfrac{u+s}{u+v+n}$. This tends to $\dfrac{s}{n}$, as the number of trials increases to infinity; and the variance of the posterior distribution tends, though more slowly, to zero.[2] Thus, like the earlier exam-

[2] See, for example, Pollard 1985, Chapter 8.

ple, the influence of the prior distribution upon the posterior distribution steadily diminishes with the size of the sample, the rate of diminution being considerable, as simple examples, which the reader may construct, would show.

Credible Intervals and Confidence Intervals

Parameter estimates are often reported in the form of a range of possibilities, e.g., $\theta = \theta^* \pm \varepsilon$. This has a natural Bayesian interpretation, namely, as a set of values possessing a high probability of containing the true value of θ. In general, if P denotes the probability that θ lies between a and b, then the interval (a, b) is said to be a $100P$ percent *credible interval* for θ. Bayesians recommend credible intervals as useful summaries of posterior distributions.

We may illustrate the idea with our first example, where we were concerned to estimate the mean of a normal population. We showed that in the circumstances hypothesized, a sufficient amount of data determined a posterior distribution with a mean equal to the sample mean, $\bar{x}$, and a standard deviation, σ_n equal to $\frac{\sigma}{\sqrt{n}}$. Since the distribution is normal, the range $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ contains θ with probability 0.95, and so constitutes a 95 percent credible interval.

Of course, there are other 95 percent credible intervals corresponding to different areas of the posterior distribution, for instance, the infinite range of values defined by $\theta > \bar{x} - 1.64\sigma_n$, and intervals that extend further into the tails of the posterior distribution while omitting a more or less narrow band of values around its centre. There is sometimes a discussion about which of these intervals should be "chosen" (for example by Lindley, 1965, Volume 2, pp. 24–25), which we believe to be misconceived, for strictly speaking, one should not choose an interval, because a choice implies a commitment in excess of that permitted by the 0.95 probability that all these intervals share. All 95 percent credible intervals are on a par from the inductive or scientific point of view.

Bayesian credible intervals resemble the confidence intervals of classical statistics. Indeed, in the particular case before us, the

95 percent credible interval depicted in the diagram and the 95 percent confidence interval that is routinely favoured (the shortest one) coincide. Yet the two types of interval are crucially different. The first states the probability, relative to the evidence, that θ lies within the interval. The second says nothing about the probability of θ, nor does it express any degree of uncertainty about θ in non-probabilistic terms, as its defenders, unsuccessfully claim that it does. Credible intervals provide an intelligible interpretation and a rational explanation for the intuitions underlying classical confidence intervals.

8.b | The Principle of Stable Estimation

The estimates delivered in our two examples are, as we showed, very insensitive to variations in the prior distributions. However, since the priors assumed in the examples were restricted, in the first case, to normal and in the second, to beta distributions, the question arises whether this insensitivity persists when these restrictions are relaxed. That it does is the burden of the Principle of Stable Estimation, due to Edwards, Lindman and Savage 1963, a practically useful aspect of a more general result proved by Blackwell and Dubins 1962.

The idea is this. Consider a parameter θ, with a prior probability density distribution $u(\theta)$, and the corresponding posterior distribution, relative to some data x, $u(\theta \mid x)$. We will denote by $w(\theta \mid x)$ the posterior distribution that would be induced if the prior were uniform. B is a credible interval based on $w(\theta \mid x)$, which is such that

$$\int_{\bar{B}} w(\theta \mid x)d\theta \leq \alpha \int_{B} w(\theta \mid x)d\theta,$$

where $\bar{B}$ is the complement of B, and α is 10^{-4} or less (that is, B is a 99.99 percent or higher credible interval). Consider secondly, the variation of the actual prior distribution within the interval B, which the second condition says should be small; more specifically, this condition stipulates that there be positive numbers φ (whose actual value is immaterial in this context) and β, the latter < 0.05, such that for all θ in B

$$\varphi \leq u(\theta) \leq (1 + \beta)\varphi.$$

Finally, consider the variation of the prior distribution outside B. The third condition stipulates that the prior should be "nowhere astronomically big compared to its nearly constant values in B". More specifically, for some positive number $\delta < 1000$ and for all θ

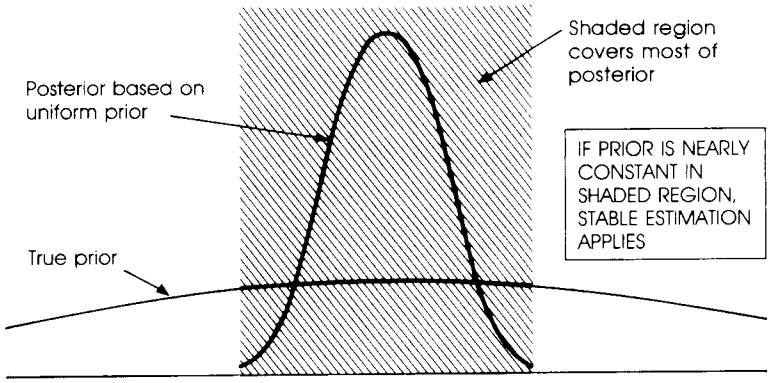
$$u(\theta) \leq \delta\varphi.$$

Edwards, Lindman, and Savage then show that under these conditions, the true posterior distribution and the one calculated on the hypothesis of a uniform prior are approximately the same. Moreover, the approximation is greater, the smaller α, β, and δ. Hence, whatever the prior beliefs of different people, so long as they meet the said conditions, the principle ensures that their posterior beliefs are all roughly the same.

In practice, the stated conditions of the Stable Estimation principle are ordinarily quite accessible, and one can check whether they hold in any particular case. This is done by first calculating the posterior distribution that would obtain if the prior were uniform,[3] then examining a 99.99 percent credible interval (the area shaded in the diagram) to see whether the range of variation of the true prior within the interval is as small as the principle requires, and finally checking outside the interval that the actual prior is never larger than the average value within it by more than the prescribed amount.

The Principle of Stable Estimation assures us that the relative independence of Bayesian estimates from the prior distributions, which we noted in our two examples, is not confined to priors belonging to particular families of distributions. The principle also tells us that provided the sample is sufficiently large, you do not need to describe the prior distribution with great accuracy or precision in order to arrive at a Bayesian estimate that *is* both

[3] Since a uniform distribution is only defined over an interval that is bounded at both ends, this will involve setting limits to the values that the parameter could have.



Practical Application of the Principle of Stable Estimation (Phillips 1973)

accurate and precise. This answers the objection that is sometimes raised against Bayesian estimation that it can rarely get off the ground because of the practical difficulties of accurately ascertaining one's own or other people's prior belief distributions.

8.c Describing the Evidence

An experimental result is a physical state; on the other hand, scientific evidence is a linguistic statement, and a question that arises is: which aspects of the physical state should go into the evidence statement? A complete description would be infinitely long and clearly impossible. Nor is it desirable even as an abstract ideal to include every aspect of an experimental result in the evidence, for some aspects are plainly irrelevant. For instance, the colour of the experimenter's shoes when sowing plant seeds would not normally be worth recording if the genetic structure of the plant were the issue. Evidence need not refer to irrelevant details such as these, but, as we showed in 5.f, it should omit no details that are relevant.

In this context, a fact is relevant to a hypothesis if knowing it affects how the hypothesis is appraised; when the method of appraisal is Bayesian, this means that the relevance of a fact is determined by whether it alters the probabilities of any of the hypotheses of interest. We may illustrate this notion of evidential

relevance with the simple, but representative problem of estimating a coin's physical probabilities of landing heads and tails, using evidence obtained from flipping it a number of times. Let us say the coin was flipped 10 times, producing 6 heads and 4 tails. The table below lists several possible descriptions of such a result and several circumstances in which it might have been obtained.

TABLE 8.1

Possible descriptions of the result *6 heads, 4 tails* obtained in various coin-tossing trials

e_1:	6 heads
e_2:	6 heads, 4 tails, in a trial designed to end after 10 tosses of the coin
e_3:	6 heads, 4 tails, in a trial designed to end after 6 heads appear
e_4:	The sequence *TTHTHHHHTH*
e_5:	The sequence *TTHTHHHHTH* obtained in a trial designed to terminate when the experimenter is called for lunch
e_6:	6 heads, 4 tails

The first description gives a very thin account of the experiment, not even revealing how many tails it produced, information that is clearly relevant, and because of this omission, its evidential value is small and hard to quantify. A Bayesian inference from the evidence statement would require not only a prior distribution over θ (the coin's physical probability of landing heads), but also one over n, the number of times the coin was tossed. It is unnecessary to enter further into this complicated case, save to note that unlike the classical approach, and more plausibly we think, the Bayesian does not necessarily find e_1 uninformative.

The second description, e_2, tells us the number of heads and tails in the result and gives the stopping rule that governed the experiment. We can use this directly in Bayes's theorem to calculate the posterior distribution $P(\theta \mid e) = \dfrac{P(\theta \mid e)}{P(e)} P(\theta)$. Since the

number of times the coin was flipped was predetermined according to the stopping rule, the likelihood function and the probability of the evidence are given in familiar fashion by the formulas:

$$P(e_2 \mid \theta) = {}^{n}C_{r}\theta^{r}(1-\theta)^{n-r} \text{ and } P(e_2) = \int_0^1 {}^{n}C_{r}\theta^{r}(1-\theta)^{n-r}P(\theta)d\theta$$

where r denotes the number of heads and n the number of coin tosses, which in our present example are 6 and 10, respectively. The binomial factor ${}^{n}C_{r}$, being a function of r and n only, is independent of θ, and so cancels out of Bayes's theorem. Clearly any other description of the experimental outcome for which $P(e_2 \mid \theta) = K\theta^{r}(1-\theta)^{n-r}$ yields the same posterior distribution, provided K is independent of θ. This is the case with e_3, which states a different stopping rule, and with e_4, which lists the precise sequence of heads and tails in the outcome. In the former case, $K = {}^{n-1}C_{r-1}$, as we showed in 5.d, and in the latter $K = 1$. Hence, the Bayesian conclusion is unaffected by whether the experimenter intended to stop after n tosses of the coin or when r heads appeared in the sample or, indeed, whether the experiment was performed without a stopping rule. As we noted in Chapter 5, this is not so in the classical scheme.

The next experimental description, e_5, illustrates a case where the rule to stop the trial depends on some external event rather than on any feature of the sample. If l states that the experiment was designed to stop as soon as lunch was ready, then e_5 is equivalent to the conjunction l & e_4. So, in applying Bayes's theorem, we need to consider the probabilities $P(l \,\&\, e_4 \mid \theta)$, which can also be expressed as $P(l \mid e_4 \,\&\, \theta)P(e_4 \mid \theta)$. If l is probabilistically independent of θ in the presence of e_4, as we have argued that it is, then $P(e_5 \mid \theta)$ has the form $K\theta^{r}(1-\theta)^{n-r}$, where K is a constant, relative to θ. And so, as we observed above, this constant cancels from Bayes's theorem, which then delivers the same posterior distribution for e_5 as for e_4, implying that the stopping rule is inductively irrelevant. It clearly should be, though the classical philosophy denies this, as we showed in Chapter 5.

Data are normally presented without mention of the stopping rule, as in e_6, which merely reports the number of heads and tails produced in the trial. This poses a problem, for the probability of such a result does depend on the stopping rule, and if $P(e \mid \theta)$

cannot be computed, neither Bayes's theorem, nor a classical test of significance can be applied. The difficulty is usually met by calculating the probabilities *as if* the sample size had been fixed in advance. In the present case, this means that $P(e \mid \theta)$ would be reckoned as equal to ${}^{n}C_{r}\theta^{r}(1-\theta)^{n-r}$. This seems arbitrary and wrong, but is in fact justified in a Bayesian (though not a classical) analysis, for the result *r heads, n – r tails* must have occurred as some particular sequence, and whatever that sequence was, its probability, conditional on θ, is $\theta^{r}(1-\theta)^{n-r}$, and we may correctly use this as the input for Bayes's theorem. But as we pointed out before, we obtain thereby the same posterior distribution for θ as we do when the calculation is based on the possibly incorrect assumption of a fixed-sample stopping rule.

We have argued that the stopping rule is irrelevant in any inductive inference, and the Bayesian endorsement of this view and the classical denial of it seem to us decisive arguments in favour of the one and against the other. But before resting our case, we should visit a couple of arguments that may appear to show that in fact the Bayesian is wrong to take no account of the stopping rule.

To illustrate the first, consider the stopping rule we mentioned earlier, in which the trial is planned to stop as soon as lunch is ready, and suppose the purpose of the trial is to use a random sample to estimate the mean height of a group of chefs who are preparing the meal while the trial is progressing. Now if tall chefs cook faster than short ones, the time taken before the random sampling stops depends on the unknown parameter and so contains relevant information about it, which should be reflected in any Bayesian analysis. However, despite a possible initial impression to the contrary, this does not endorse the classical position concerning the stopping rule, for it does not imply that the experimenter's subjective intention or plan to stop the trial at a particular point had any inductive significance; all that was relevant in this contrived and unusual case was the objective coincidence of lunch being ready at the same moment as the experiment stopped.

A second argument that is made from time to time is this: a Bayesian who fails to announce in advance the circumstances under which a trial will be stopped could decide to continue to sample for as long as it takes to reach any desired posterior probability for any hypothesis, however remote from the truth it might be. And according to Mayo (1996, pp. 352-7), for example, the determined Bayesian will assuredly succeed in this eventually, that is, in the limit as the sampling is repeated to infinity, or, as Mayo puts it, by "going on long enough".[4] Mayo suggests that this putative property allows the Bayesian to "mislead", and vitiates any Bayesian estimate.

But in fact, neither Mayo nor anyone else has demonstrated that Bayesian conclusions drawn from the results of such try-and-try-again experiments would be misleading, or spurious, or in any way wrong. And unless such demonstrations are forthcoming, there is no case whatever for the Bayesian to answer.

Moreover, the main premise of the objection is wrong, for, as Savage (1962) proved, the application of Bayes's theorem does not guarantee supportive evidence for any hypothesis, however long the sampling is continued, and in fact, as Kadane *et al.* (1999, Chapters 3.7 and 3.8) have shown, the probability of eventually obtaining strongly supportive evidence for a false theory is small. How small this probability is depends on the prior probability of the hypothesis and on the degree of confirmation sought for it.

Sufficiency

The Bayesian concept of relevant information is that E is relevant to θ if and only if the posterior distribution of θ given E is different from its prior distribution. Of the items of evidence e_1 to e_6 in the above table, the last contains the least amount of information and yet, as we showed, it is just as informative about θ as the others. So the information about the various stopping rules in those pieces of evidence is, in a Bayesian sense, inductively irrelevant. Classical statistics employs the idea of relevant information too, approaching it through its concept of sufficient statistics. It will be recalled from Chapter 5 that a statistic t is defined as *sufficient for* θ, relative to data x, when $P(x \mid t)$ is independent of θ, this being interpreted as t containing all the information in x that is

[4] Of course, in practical reality, the sampling could never be continued very long at all, let alone be taken to the limit.

relevant to θ. Here, x and t are random variables, and the sufficiency definition refers to all the possible values they take in the experimental outcome space. Since the latter is determined by the stopping rule, the classical requirement that inferential statistics be sufficient underlines the centrality of the stopping rule in classical inference.

We argued that e_4: the sequence *TTHTHHHHTH* and e_6: *6 heads, 4 tails* contain the same information about θ. Yet if the number of coin tosses were 10,000, rather than 10, the same argument might not hold. For if some pattern were detected in the sequence, for example, if all the heads preceded all the tails, or if the first half of the sequence showed a large preponderance of tails and the second half of heads, there would be evidence that the coin's centre of gravity had shifted, which would be missed if one relied simply on the number of heads and tails. The lesson here is that the scientist should examine evidence in as much detail as is feasible, so as not to overlook significant information. The possibility that the order of the heads and tails might be evidentially significant did not emerge in our earlier discussion, because we, in effect, treated such variant hypotheses as having zero probability. But this was a simplification and in practice, the open-minded scientist would not usually take this most extreme attitude to unlikely hypotheses.

8.d Sampling

In order to derive a posterior distribution for a parameter from sample data, one needs to compute the likelihood terms that figure in Bayes's theorem. Likelihoods are also required in classical inferences. And since these inferences are supposed to be objective, the likelihoods they employ must also be, to which end classical statisticians generally call for experimental samples to be created by means of a physical randomizing device, in order to ensure that every element is endowed with precisely the same objective probability of being included in the sample. This means that a sample drawn haphazardly, or purposively, in the manner we described in 5.g, would be uninformative and unacceptable for the purpose of a classical estimation, while the very same sample,

had it been generated by a random process, would have been perfectly alright. This contrast sounds paradoxical, and indeed, Stuart (1962, p. 12) described it as a "paradox of sampling" and as a bitter pill that must be swallowed, in the interests of what he regarded as the only valid estimation methods, namely those of classical inference.

In fact, Bayesian inference is also sensitive to the sampling method, and as we show, there is nothing paradoxical about this, since the way that the data were collected may indeed carry useful information. Suppose, for example, that in order to estimate the proportion θ of As in a population, a sample of 1 (to cut the argument to the bare bones) was drawn and that this was an A; and suppose that the selected element also possesses some other characteristic, B. The goal might be to discover what proportion of the population intends to vote for a particular political party, and the sample might consist of an individual who does so intend, and who is noted to be above the age of 60. The posterior distribution in the light of this information is given by

$$\frac{P(\theta \mid AB)}{P(\theta)} = \frac{P(AB \mid \theta)}{P(AB)} = \frac{P(A \mid \theta B)}{P(A \mid B)} \times \frac{P(B \mid \theta)}{P(B)} .$$

The sample could have been gathered in a number of ways. For example, it might be the result of a random draw from the entire population, or alternatively, from just that portion of the population containing Bs. If the latter, then $P(B) = P(B \mid \theta) = 1$. If the former, the equality between $P(B)$ and $P(B \mid \theta)$ holds only if B and θ are probabilistically independent, in which event, the two sampling methods lead to the same Bayesian conclusion. But this is a particular case, and as a general rule the inductive force of a given sample is not independent of the selection procedure. (See 5.d above, and Korb 1994.)

This is intuitively right, we suggest, because the number of Bs contained in a random sample that was collected with the purpose of estimating θ is also a measure of the overall proportion of Bs. If the latter is probabilistically dependent on θ, it will convey some information about that parameter too. On the other hand, if you deliberately restricted the sample to Bs, the fact that it contains only elements of that type would carry practically no

information about their frequency in the population; hence, that potential source of knowledge about θ would be unavailable.

Bayesian and classical positions agree then that the sampling method as well as the sample is inductively relevant. But they do not agree on the role of random samples. The classical position is that only random samples that are created using a physical randomizing mechanism can be informative, making the inconvenient demand that other sorts of sample should not be used. Bayesian induction, on the other hand, can operate satisfactorily using what we described in Chapter 5 as purposive or judgment sampling.

8.e | Testing Causal Hypotheses

The most widely used statistical methodology for testing and evaluating causal hypotheses, particularly in medical and agricultural trials, was invented by Fisher and is rooted in classical procedures of inference. It will be recalled from our discussion in Chapter 6 that the novelty of Fisher's approach was to require a process known as randomization. We have argued that despite the weight of opinion that regards it as a *sine qua non*, the randomizing of treatments in a trial does not do the job expected of it and, moreover, that in the medical context, it can be unnecessary, inconvenient, and unethical. We present here (following Urbach 1993) an outline of a Bayesian way of testing causal hypotheses, which is soundly based and, we believe, intuitively more satisfactory than the Fisherian method.

Consider the matter through a medical example. Suppose a new drug were discovered which, because of its structural similarity to an established cure for depression, seems likely also to be an effective treatment for that condition. Or suppose that the drug had given encouraging results in a pilot study involving a small number of patients. (As we said earlier, without some indication that the drug is likely to be effective, a large-scale trial of an experimental drug or treatment is indefensible, either economically or ethically.) A trial to test the drug's efficacy would normally take the following form. Two groups of sufferers would be constituted. One of them, the test group, would receive the drug, while

the other, the control group would not. In practice, the experiment would be more elaborate than this; for example, the control group would receive a placebo, that is, a substance which appears indistinguishable from the test drug but lacks any relevant pharmacologically activity; and patients would not know whether they were part of the drug group or the placebo group. Moreover, a fastidiously conducted trial would ensure that the doctor too is unaware of whether he or she is administering the drug or the placebo (trials performed with this restriction are said to be *double blind*). Further precautions might also be taken to ensure that any other factors thought likely by medical experts to influence recovery were equally represented in each of the groups. It is then said that these factors have been controlled for.

Why such a complicated experiment? Well, the reason for a comparison, or control group is obvious. We are interested in the causal effect of the drug on the chance of recovery; so we need to know not only how patients react when they are given the drug, but also how they would respond in its absence, and the conditions in the comparison group are intended to simulate the latter circumstance. The requirement to match or control the groups, in certain respects, is also intuitive; but although always insisted upon, it is never derived from epistemological principles in the standard, classical expositions. However, selective controls in clinical trials do have a rational basis. It is provided by Bayes's theorem, as we shall now explain.

Clinical Trials: a Bayesian Analysis

To simplify our exposition, we will consider a particular trial in which, for illustration, 80 percent of a specified number of test-group patients have recovered from the disease in question, while the recovery rate in the control group is 40 percent: call this the evidence, e. Ideally, we would like to be able to conclude that these percentages also approximate the probabilities of recovery for similar people outside the trial. This hypothesis, which is represented below as H_{α}, says that the physical probability of recovery (R) for a person who has received the drug is around 0.80, and for someone who has not received the drug it is around 0.40,

provided they also satisfy certain conditions, L, M and N, say. These conditions, or what we earlier called prognostic factors, might, for example, specify that the patient's age falls in a certain range, that he or she has reached a certain stage of the illness, and so forth. (In the formulations below, *Drug* and *~Drug* denote the conditions of the drug's presence and absence, respectively.) For H_α to explain e, we also need to be able to assert that the conditions L, M and N were satisfied by the subjects in both of the experimental groups, and that the test group received the drug, while the control group did not; this is the content of H_β. So the hypothesis claiming that the drug caused the observed discrepancy in recovery rates is the combination H_α & H_β, which we label H and call the drug hypothesis:

H_α: $P(R \mid L, M, N, Drug) \approx 0.80$ & $P(R \mid L, M, N, \sim Drug) \approx 0.40$

(H)

H_β: Patients in the experimental groups satisfy conditions L, M and N.

But the drug hypothesis is not the only one capable of explaining the experimental findings. Another explanation might attribute them to a psychosomatic effect induced by a greater level of optimism amongst the test group patients than amongst those in the control group. Let H'_α signify the hypothesis that under conditions L, M and N, the drug has no effect on the course of the disease, but that an optimistic attitude (O) promotes recovery. By parallel reasoning, the hypothesis that explains the evidence as a psychosomatic, confidence effect is the combination of H'_α and H'_β, which we label H':

H'_α: $P(R \mid L, M, N, O) \approx 0.80$ & $P(R \mid L, M, N, \sim O) \approx 0.40$

(H')

H'_β: Test group patients satisfy L, M, N, O; control group patients satisfy L, M, N, $\sim O$.

The Bayesian method addresses itself to the probabilities of hypotheses and to how those probabilities are updated via Bayes's theorem in the light of new information. Let us see how this updating works in the present case. To start with, and for the purpose of exposition, we shall suppose that H and H' are the only hypotheses with any chance of being true. In that event, Bayes's theorem takes the following form:

$$P(H \mid e) = \frac{1}{1 + \dfrac{P(e \mid H')P(H')}{P(e \mid H)P(H)}}.$$

The question is how to design the experiment so as to maximize the probability of the drug hypothesis for some given e. We see from the above equation that the only components that can be manipulated to this end are $P(H)$ and $P(H')$, and that the smaller the latter and the larger the former, the better. Now, if the two components of the hypotheses are independent—a reasonable premise, which also simplifies our argument—then $P(H) = P(H_\alpha)P(H_\beta)$ and $P(H') = P(H'_\alpha)P(H'_\beta)$. The goal of minimizing $P(H')$ now comes down to that of minimizing $P(H'_\beta)$. And again, since only $P(H_\beta)$ can be changed by adjusting the experimental conditions, it is this component of $P(H)$ that should be maximized.

Now H'_β states, amongst other things, that patients in the test group were confident of recovery, while those in the control group were not. We can reduce the probability of this being the case—reduce $P(H'_\beta)$, that is—by applying a placebo to the control group; for if the placebo is well designed, patients will have no idea which experimental group they were in, so that a number of factors that would otherwise create different expectations of recovery in the two groups would be absent. In many cases, the probability of H'_β could be reduced further by ensuring that even the doctors involved in the trial cannot distinguish the treatment from the placebo; such trials are, as we said earlier, called 'double blind'. By diminishing $P(H'_\beta)$ in these various ways, the probability of H, for a given e, is increased.

The drug hypothesis would, as we pointed out, also be made more probable by adopting appropriate measures to raise $P(H_\beta)$, that is to say, measures that increase the chance that the factors

which the drug hypothesis says are relevant to recovery (namely, *L, M* and *N)* are represented equally in the two groups. Now suppose one of those factors was virulence of the disease, or the strength of the patient, or the like, which may be hard to measure but which doctors may well be able, through long experience, to intuit. And suppose we left it to these doctors to construct the comparison groups. It would not be surprising if the resulting groups contained unequal numbers of the more vulnerable patients. For doctors should be and generally are guided by the wish to secure the best treatment for their patients, in accordance with the Hippocratic Oath,[5] and if they also entertained prior views on the trial treatment's efficacy, they would be inclined to distribute patients among the trial groups according to particular medical needs, rather than with impartiality. And there may be other unconscious or even conscious motivations amongst experimenters that could lead them to create test groups with an inbuilt bias either in favour or against the test treatment.[6] This is where randomized allocation can play a role, for it is a mechanism that excludes the doctor's feelings and thoughts from the process of forming the experimental groups and thereby makes balanced groups more probable.

We have simplified this exposition by considering a single alternative to the drug hypothesis. There will normally be many alternatives, each contributing an element to the denominator, which would then include as summands other terms of the form: $P(e \mid H'')P(H'') = P(e \mid H''_{\alpha}\&H''_{\beta})P(H''_{\alpha})P(H''_{\beta})$—again, assuming independence. Each of these terms relates to specific, hypothetical prognostic factors, and each needs to be minimized in order to maximize the posterior probability of H, the drug hypothesis; as we explained, this is done is by matching the experimental groups in ways that are suited to reducing $P(H''_{\beta})$.

But not every conceivable factor can be matched across the groups; nor is a comprehensive matching in any sense either needed or desirable. For, consider a possible prognostic factor whose causal influence is described in the hypothesis H''_{α} and suppose that hypothesis to be very improbable; in that event, the whole of the corresponding term in Bayes's theorem would already be extremely small, and so the advantage of reducing it further by applying appropriate controls would be correspondingly small. And in most cases, that small advantage would be outweighed by the extra cost and inconvenience of the more elaborate trial. For example, though one could introduce a control to reduce the probability that the clinicians treating the two groups had different average shoe sizes, such a precaution would only negligibly affect the posterior probability of the drug hypothesis, since shoe size is so immensely unlikely to influence the outcome of the trial.

In summary, the Bayesian theory explains why the experimental groups in clinical trials need to be matched in certain respects, and why they need not be matched in every respect. It agrees with common sense in affirming that *the chief concern when designing a clinical trial should be to make it unlikely that the experimental groups differ on factors that are likely to affect the outcome.* Designs that achieve such a balance between groups are termed "haphazard" by Lindley, 1982a, p. 439, though we prefer to think of them as adequately matched or controlled.

With this rule in mind, it is evident that a randomized allocation of subjects to treatments might sometimes be useful in clinical trials as a way of better balancing the experimental groups, insofar as it stops those making the allocation from doing so on the basis, for example, of how sick the patient is. *But randomized allocation is not absolutely necessary; it is no* sine qua non*; it is not the only or even always the best way of constructing the treatment groups in a clinical trial.*

Clinical Trials without Randomization

Bayesian and classical prescriptions for clinical trials differ in two respects of practical importance. First, a Bayesian analysis per-

[5] Hippocrates (died 380 360 B.C.) was the author of the eponymous oath by which doctors committed themselves to certain professional standards. Graduates in some medical schools today formally take a modernized version of that oath. The classical formulation enjoins doctors, amongst other things, to keep their patients "from harm and injustice". A form that is currently approved by the American Medical Association makes the doctor promise "that you will exercise your profession solely for the cure of your patients".

[6] Kadane and Seidenfeld (1990) cite the possibilities of bias arising because the experimenter was the inventor of the test treatment, with a personal stake in its success, or was keen to make a splash in the academic literature by announcing a surprising result.

mits the continuous evaluation of results as they accumulate, thus allowing a trial to be halted as soon as the effectiveness or otherwise of the experimental treatment becomes apparent. It will be recalled that, by contrast, when classical principles are strictly observed, a clinical trial that has been brought to an unscheduled stop cannot even be interpreted, whatever the results recorded at that stage. And sequential clinical trials, which were specially invented in order to allow interim analyses of trial results using the classical techniques are ineffective, as we argued in Chapter 6.

The second difference relates to the formation of the comparison groups in a clinical trial. On the Bayesian view, the essential desideratum is that the groups be adequately matched on likely prognostic factors. Freed from the absolute need to randomize, Bayesian principles affirm, what its rival denies, that decisive information about a medical treatment can be obtained from trials that use, for example, a historical control group, that is to say, a number of patients who have already received an alternative therapy or no therapy at all. Such groups may be constructed by means of information derived from medical records and, provided they are adequately matched on prognostic factors with a test group (which may also be of the historical type), an informative comparison is allowed.

Historically controlled trials are widely dismissed as inherently fallacious by classically minded medical statisticians, who regard concurrent, randomized trials as the "only way to assess new drugs" (McIntyre 1991). But this view is not only indefensible on epistemic grounds, it is daily refuted in medical practice, where doctors' knowledge and expertise is accumulated in part by continual, informal comparisons of current and former patients.

Even some classical statisticians (such as Byar *et al.* 1990) concede that historical comparison groups may be preferable in certain circumstances, particularly where a randomized trial would expose critically ill patients in a control group to a useless placebo, while others are given a promising, experimental treatment. Byar and his colleagues claimed that the standard trial structure should be suspended when there is "a justifiable expectation that the potential benefit to the patient will be sufficiently

large to make interpretation of the results of a non-randomized trial unambiguous". But they do not say what process of inference could deliver the unambiguous interpretation of results. No classical significance test can be intended, since Byar makes the familiar classical claim, which we discussed in Chapter 6, that "it is the process of randomization that generates the significance test". Nor does any other tool of classical inference seem to be available. It is more likely that Byar *et al.* are here allowing themselves an informal Bayesian interpretation, for they say that a historically controlled trial is acceptable only if the trial treatment has a strong prior credibility: "the scientific rationale for the [trial] treatment must be sufficiently strong that a positive result would be widely expected".

The admissibility in principle of historically controlled trials is not just an interesting theoretical implication of the Bayesian view, but is of considerable practical significance too. Firstly, such trials call for fewer new subjects, which is of particular importance when rare medical conditions are the subject of study. And smaller trials are generally cheaper. Secondly, historical comparison groups do not expose subjects to ineffective placebos or to what clinicians expect will turn out to be inferior comparison treatments, considerations that address natural ethical concerns, and mitigate the reluctance commonly found amongst patients to participate in trials.

Historically controlled trials are however, not easy to set up. The comparison groups can only be formed with the aid of thorough medical records of past patients, more detailed than the records that are routinely kept, and more accessible. To this end, Berry 1989 proposed the establishment of national databases containing patients' characteristics, diagnoses, treatments and outcomes, which could be open to the public and contributed to by every doctor. Some modest work along these lines has been done. Unfortunately, the widely held, though erroneous opinion that historical controls are intrinsically unacceptable or impossible has discouraged efforts to overcome the purely practical obstacles that stand in the way of effective historically controlled trials. Here is a case where the mistaken principles of classical methodology are harmful.

Summary

Bayes's theorem supplies coherent and intuitive guidelines for clinical and similar trials, which contrast significantly with classical ones. One striking difference between the two approaches is that the second simply takes the need for controls for granted, while the first explains that need and, moreover, distinguishes in a plausible way between factors that have to be controlled and those that do not. Another difference is that the Bayesian approach does not make the random allocation of subjects to treatments a universal, absolute requirement. We regard this as a considerable merit, since we have discovered no good reason for regarding a random allocation as indispensable and several good reasons for not so regarding it.

8.f | Conclusion

The Bayesian way of estimating parameters associates different degrees of confidence with different values and ranges of values, as classical statisticians sought to do, though as we have seen, they were unsuccessful in this. It makes such estimates through a single principle, Bayes's theorem, which applies to all inferential problems. Hence, the Bayesian treatment of statistical and deterministic theories is the same and is underwritten by the same philosophical idea.

Bayesian estimation also chimes in well with our intuitions. It accounts for the intuitively plausible sufficiency condition and for the natural preference for maximum precision in estimators, while finding no place for the criteria of 'unbiasedness' and 'consistency', which, we have argued, are based on error. The Bayesian method also avoids a perverse feature of classical methods, namely, a dependence on the outcome space and hence on the subjective stopping rule.

There is a subjective element in the Bayesian approach which offends some, but which, we submit, is wholly realistic. Perfectly sane scientists with access to the same information often do evaluate theories differently. Newton and Leibniz differed sharply on gravitational theory; Einstein's opinion of Quantum theory in its

Copenhagen interpretation was at variance with that of most of his colleagues; for many years, distinguished astronomers defended the Big Bang theory, while equally distinguished colleagues defended with equal vigour the competing Steady State theory, and similar divergences of view continue to occur in every branch of physical science, in medicine, biology and psychology. Bayesian theory anticipates that many such divergences will be resolved as probabilities are revised through new evidence, but it also allows for the possibility of people whose predispositions either for or against certain theories are so pronounced and distinct from the norm that their opinions remain eccentric, even after a large mass of relevant data has accumulated. You might take the view that such eccentrics are pathological, that every theory has a single value relative to a given body of knowledge, and that responsible scientists ought not to let personal, subjective factors influence their beliefs. But then you would have to face the fact that this view is itself a prejudice, because despite an immense intellectual effort, no one has produced a coherent defence of it, let alone a proof.

After decades when Bayesian ideas and methods were despised and rejected, their merits are now coming to be widely acknowledged, even by orthodox statisticians, so much so that many are willing to put Bayesianism and Frequentistism on a par. But this is not an outcome that we regard as satisfactory. Our arguments have been that the first is well founded and the second not, and that the second should give way to the first. Blasco (2001, p. 2042) provides an example of the even-handedness that we reject. Writing on "The Bayesian controversy in animal breeding", he argues that in practice there is no need to take sides:

> If the animal breeder is not interested in the philosophical problems associated with induction, but in tools to solve problems, both Bayesian and frequentist schools of inference are well established and it is not necessary to justify why one or the other school is preferred. Neither of them now has operational difficulties, with the exception of some complex cases. . . . To choose one school or the other should be related to whether there are solutions in one school that the other does not offer, to how easily the problems are solved, and to how comfortable the scientist feels with the particular way of expressing the results.

But we have argued that what Blasco calls "the philosophical problems associated with induction" ought to be of interest to the practical scientist, for although frequentist and Bayesian tools give superficially similar results, and recommend superficially similar trials in some cases, they also make crucially different recommendations in others. We have shown that frequentist tools do not solve *any* problems. The conclusions they license ("the best estimate of the unknown parameter, θ, is such and such"; "such and such is a 99 percent confidence interval for θ"; "h_0 is rejected at the 5 percent level", etc.) have no inductive significance whatever. True, one can often draw frequentist conclusions "easily", but this is of no account and does not render them scientifically meaningful. True, many scientists feel "comfortable" with frequentist results, but this, we suggest, is because they are misinterpreting them and endowing them with a meaning they cannot possibly bear.

CHAPTER 9

Finale: Some General Issues

9.a The Charge of Subjectivism

The theory of inductive inference, or inductive *logic* as we feel entitled to call it, that we have presented here is sometimes called the subjective Bayesian theory, 'subjective' primarily because it imposes no constraints on the form of the prior probabilities in Bayes's Theorem calculations. Many both inside and outside the Bayesian camp think this results in a theory inadequate to its presumed purpose of furnishing an objective account of inductive inference, since these priors will necessarily have to be 'subjective', and subjectivity should have no place in scientific inference.

Those who have read the earlier chapters will recall that our view is that all this is incorrect. Firstly, some subjective element exists in all scientific appraisal and it is a merit of this theory that where it occurs it is signalled explicitly, not concealed from view. Secondly, what we have in this theory is in fact a perfectly objective logic of inductive inference, whose 'premises' can be just those prior probabilities, with Bayes's Theorem as the inference engine generating a valid conclusion: the posterior distribution. As we pointed out, the situation is wholly analogous to deductive logic, where the logic is the inference engine: you choose the premises, and the engine generates the valid conclusions from them. De Finetti characteristically puts the position very well:

> We strive to make judgments as dispassionate, reflective and wise as possible by a doctrine that shows where and how they intervene and lays bare possible inconsistencies between judgments. There is an instructive analogy between [deductive] logic, which convinces one that acceptance of some subjective opinions as 'certain' entails the certainty of others, and the theory of subjective probabilities, which similarly connects uncertain opinions. (1972, p. 144)

Nevertheless, it cannot be proved that there are no further acceptable constraints which we have simply missed or ignored which might substantially reduce or even in suitable cases eliminate the degrees of freedom in the selection of prior distributions. Indeed, the history of the Bayesian theory is to a considerable extent the history of attempts to find such constraints, and we shall end by looking at the principal ones.

9.a.1 The Principle of Indifference

A constraint additional to those we presented in Chapter 3, and defined consistency with respect to, has been of great historical importance, and even today we still find its claim to inclusion strongly pressed. Called by Keynes the *Principle of Indifference*, it is a *prima facie* highly plausible symmetry principle enjoining that a symmetric relationship between the members of a partition should be respected by a correspondingly symmetric *a priori* distribution of probabilities. More precisely: *equal parts of the possibility space should receive equal probabilities relative to a null state of background information.*

Not only did this principle, and to some people it still does, seem intuitively compelling: it also turned out to have impressive methodological power (as we shall see, more so even than its advocates immediately realised). Thomas Bayes was the first to use it to prove a deep and important result in statistics, the so-called 'inversion' of Bernoulli's Theorem. Recall from Chapter 2 that James Bernoulli had proved a fundamental theorem about, in modern notation, a possibility space Ω consisting of n-fold sequences s of 0s and 1s, a class of events defined in Ω including all events of the form 'there is a 1 at the ith index' (also describable in the language of random variables by a formula 'X_i=1', where the X_i are n $\{0,1\}$-valued random variables defined on Ω), and a probability function assigning objective probabilities to these events such that '$X_i = 1$' has the same probability p independently of i, and the events '$X_1 = x_1$', ... ,'$X_n = x_n$', $x_i = 0$ or 1, are independent. Bernoulli's result was that for any small $\varepsilon > 0$, the probability that the absolute value of the difference between the relative frequency $(n^{-1})\Sigma X_i$ of 1s (up to n) and p is less than ε tends to 1 as n tends to infinity.

People, including Bernoulli himself, wanted to interpret this result as licensing an inference that the probability is very large, for large enough n, that p is close to the observed relative frequency. Unfortunately, no such inference is licensed. Bernoulli's Theorem is a result about a probability distribution characterised by a real-valued parameter p: hence p is not itself not a random variable to which probabilites are attached. To 'invert' Bernoulli's theorem requires defining a more extensive possibility space Ω', a class C of propositions defined in Ω', and a probability function P defined on C, in which '$p = r \pm \delta$' is an event in C, i.e. in which p is a random variable.

This is just what Bayes attempted to do, and to a great extent succeeded in doing, and thereby came up with the first rigorously obtained posterior distribution for a statistical hypothesis, in this case about the value of p, now explicitly considered as a random variable. Here is a brief sketch, in more modern notation than Bayes used, of how to perform the calculation. In fact, it is a fairly straightforward application of Bayes's Theorem, in which the likelihood, i.e. the probability of the data, r 1s and $n - r$ 0s, conditional on p, is given by the function ${}^nC_r\, p^r(1-p)^{n-r}$, and the prior probability of p determined by the Principle of Indifference as the uniform distribution with density 1 over the unit interval [0,1] (in other words, it is simply assumed that the prior probability of the Bernoulli model, of constant probability and independence, is 1; Bayes's own example was carefully crafted so that the assumption is not made explicit). Thus the conditional probability density is, by Bayes's Theorem, proportional to

$$f(p \mid r) = {}^nC_r\, p^r(1-p)^{n-r}$$

But since $f(p \mid r)$ is a continuous probability density it must integrate to 1. Hence we must have

$$P(X_i = r) = {}^nC_r \int_0^1 p^r(1-p)^{n-r}dp$$

and

$$f(p \mid r) = \frac{p^r(1-p)^{n-r}}{\int_0^1 p^r(1-p)^{n-r}dp}$$

We have been here before (in Chapter 3). This density is a *beta-density* with parameters r, $n - r$ (see Chapter 8), whose mean is $(r + 1) / (n + 2)$, which obviously tends to r/n. The variance is $(r + 1)(n - r + 1) / (n + 2)^2(n + 3)$, which tends to 0 as n increases, and so the posterior probability of p lying in an arbitrarily small interval around r / n tends to 1. Bayes seemed to have done what Bernoulli's Theorem could not do: tell us that it is overwhelmingly probable that in a large sample p, now explicitly a random variable over the enlarged possibility-space including not only the possible outcomes of the X_i but also the possible values of the binomial probability, is approximately equal to the observed relative frequency. As a matter of historical accuracy we should note that this was not Bayes's own method of derivation: he used a geometrical construction, in the manner of Newton's *Principia*, instead of the analytical calculation (Laplace was the first to use the integral formulas).

Of course, Bayes's posterior distribution is itself implictly conditional upon another assumption, namely that the process is a Bernoulli process, but since Bernoulli himself assumed much the same thing it is unfair to charge Bayes with making additional modelling assumptions. The important innovation here is the use of the Principle of Indifference to determine the prior distribution of p, though even the Principle itself was one that in a general way Bernoulli also endorsed, under the name of the *Principle of Insufficient Reason*. At any rate, in this example and in any others in which the hypothesis-parameter takes values in a bounded interval, the Principle of Indifference appears to yield a fully determinate posterior distribution. often something else too. Continue with the Bayes example and consider the probability $P(X_{n+1} = 1 / \Sigma X_i = r)$, where the sum is from 1 to n. By the rule for conditional probabilities this is equal to the ratio $P(X_i = r + 1) / P(\Sigma X_i = r)$, where the first sum is to $n + 1$ and the second to n. Using the reasoning above we can see that this ratio is equal to

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

which is equal to $(r + 1)/(n + 2)$. (1) is historically famous; it is known as the *Rule of Succession*. Clearly, as n grows large, the conditional probability that the next observation will reveal a 1 given that there have been r in the previous n observations tends to the observed relative frequency r / n. Thus, not only should you adjust your degree of belief in a head falling at the next toss to the observed relative frequency of heads; this result tells you exactly how you should adjust it.

The Rule of Succession was also put to less mundane uses. Keynes, in a memorable passage, wrote that

> no other formula in the alchemy of logic [*logic*, note] has exerted more astonishing powers. For it has established the existence of God from total ignorance, and it has measured with numerical precision the probability that the sun will rise tomorrow. (Keynes 1921, p. 89)

Indeed, it was Laplace himself who used the Rule to compute the odds on the sun's rising the next day, given that it had risen 1,826,213 days (by Laplace's calculation) days, to be 1,826,214 to one. Soon the Rule was seen as a justification for general enumerative induction, adjusting your confidence in any future event according to the frequency with which you have observed it in the past. The problem of induction, which Hume had declared with the most telling arguments (see Howson 2000 for an extended discussion) to be insoluble, seemed half a century later to be solved. Hume had argued that all 'probable arguments' from past observations to future predictions are necessarily circular, presupposing what they set out to prove. In appearing to show that all that was required was, on the contrary, a complete lack of commitment between the possible alternatives, the Principle of Indifference looked like the answer to Humean scepticism. Indeed, this is exactly how Bayes's result was seen by his executor and friend Richard Price, who argued in his introduction to Bayes's *Memoir* that it contained the answer to Hume's sceptical arguments.

Nemesis took some time to arrive, but arrive it eventually did, in the form of an increasing level of criticism and simple recipes for generating apparently paradoxical results from the Principle. Here is a modern example. It is expressed in the context of formalised logical language. The purpose of this is to show that the underlying problem is not, as it is often held to be (this standard defence was first resorted to by Keynes), the fault of imprecise definition. Consider two languages L_1 and L_2 in the usual modern

logical symbolism, with the identity symbol = included, as is usual, as a logical constant in both languages. Suppose L_1 and L_2 possess just one predicate symbol Q, the difference between the two languages being that L_1 has two individual names, symbolise them ***a*** and ***b***, while L_2 has none. Now it is a remarkable fact that the propositions

S_1: There is at least one individual having Q

S_2: There are exactly two individuals

can be formulated *by exactly the same formal sentences in each language*. At their simplest, these are

S'_1: $\exists x Q(x)$

S'_2: $\exists x \exists y[(x \neq y) \ \& \ \forall z(z = x \ \mathbf{v} \ z = y)]$.

Suppose our background information is conveyed by S'_2. If we stipulate that ***a*** and ***b*** name distinct individuals there are four L_1-atomic possibilities given S'_2, *viz* both have Q, ***a*** has Q and ***b*** doesn't, ***b*** has Q and ***a*** doesn't, neither has Q). In L_2 there are just three, since with respect to L_2 the individuals are indistinguishable. The Principle of Indifference would seem to demand that, conditional on S'_2, the probability of S'_1 is 3/4 if we are talking in L_1, but 2/3 if we are talking in L_2 (the stipulation that ***a*** and ***b*** name distinct individuals is there for simplicity of calculation: the non-identity of the probabilities persists, though the precise values differ from this example, if it is dropped).

The standard defence against this sort of example is that since L_1 makes more refined discriminations, one should choose that and not the apparently coarser-grained L_2 as the basis of any application of the Principle of Indifference. The defence does not, however, succeed, for a number of reasons. Firstly, and perhaps surprisingly, *it is simply not true that a language with individual names necessarily makes finer distinctions than one without.* Modern physics famously supplies striking counterexamples. Thus paired bosons—particles with spin equal to one in natural units—are indistinguishable according to quantum mechanics; this means that the coupled system (tensor product) of two such

particles has only three eigenstates of spin, corresponding to both particles having spin up (in a given direction), both having spin down, and one having spin up and the other spin down, giving exactly the same quantum statistics as L_2. (Sudbery 1986, pp. 70–74; the quantum mechanical predictions for bosons and fermions, which have even more dramatic statistics, are highly confirmed, raising the problem of how particles can in principle seem to be incapable of individuation, while at the same time remaining a plurality. Fortunately, this is a puzzle for the metaphysics of physics, not for us.) Less exotically, pounds in bank accounts are also indistinguishable in the same way: it makes no sense to say that there are four distinct ways in which two pounds can be distributed between two accounts. In these sorts of cases, therefore, there simply is no finer-grained reality than that described by L_1.

Secondly, the objection begs the question: why *should* one choose the finer partition? We are interested in seeing Bayesian probability in a logical perspective; indeed, we feel that it is only in such a perspective that it makes sense (where, of course, it makes a lot of sense). But logic does not dictate that one should choose the finer partition. Logic does not dictate that one ought to do anything at all, for that matter. But certainly not that.

Thirdly, and devastatingly, the Principle of Indifference gives different answers for *equally fine* partitions. It asserts that equal parts of a possibility-space Ω should receive the same a priori probability in the absence of any discriminating information. Here there is implicit reference to a *metric* according to which parts of Ω are judged equal. Where Ω is a continuum the metric is induced by a mapping onto an interval of real numbers, with its standard Euclidean metric (if Ω is itself an interval of real numbers the identity map is one such map). Such spaces are of course typical of those studied in mathematical statistics, but they have the property that the choice of metric can be made in different, but equivalent, ways. In fact, there is an infinity of different mappings of Ω into the real numbers such that the induced distance between two points in Ω depends on which mapping is employed. Thus equality of distance, which the Principle of Indifference exploits in continuous spaces, is strongly non-invariant.

Consider, for example, the space [0,1] of possible values of p that we considered in connection with Bayes's famous result. The

mapping $p' = p^2$ is a continuous differentiable bijection of [0,1] onto itself. The Principle of Indifference applied to the p-representation tells us that there is a probability of 1/2 of p not exceeding 1/2. But since p is $\sqrt{p'}$ the probability of p not exceeding 1/2 is, by the probability calculus, the same as the probability of p' not exceeding 1/4, and so that must be 1/2. But the Principle of Indifference applied to p' tells us that the probability that p' does not exceed 1/4 is 1/4. So we have a contradiction, and a real one this time. The objection that there is a unique 'natural' representation in the set of real numbers can be dismissed: firstly, it begs the question, and secondly, it is not even borne out in the ordinary practice of science. As Jeffreys points out, it is not uncommon for different representations to be used in different contexts; for example, some methods of measuring the charge on the electron give e, others e^2 (1961, p. 120).

As a postscript to this discussion of the Principle of Indifference we should recall that for large samples the form of the prior distribution is not of particular importance as long as it is not too extreme: as we saw earlier, for large random samples the posterior distribution for a parameter will typically converge within a small interval of its maximum-likelihood estimate independently of the exact form of the prior, so long as the latter is not too extreme (a proof of this result is given in Lindley 1965, pp. 128, 129). Thus the posterior distribution for Bayes's binomial parameter p will be concentrated within a small interval of the observed relative frequency (the maximum-likelihood estimator of p) even without assuming the uniform prior distribution. Whether the anxiety that Bayes is known to have felt about using the Principle of Indifference (though he did not call it that) might have been mitigated if he had known of this result is not known, but it is doubtful: we are not allowed very often the luxury of large amounts of data, and for most problems the form of the prior will affect the posterior. Bayes wanted a prior that reflected an even-handed ignorance, whatever the sample size, and that problem, if it is a problem—we shall argue that it is not—is not solved, though it might on occasion be mitigated, by asymptotic considerations.

9.a.2 Invariance Considerations

The Principle of Indifference is a symmetry principle stating that logical symmetries should be reflected, in the absence of any discriminating information, in uniform *a priori* probability distributions. The trouble, as we have seen, is that in continuous spaces there are too many symmetries for any one uniform distribution to reflect. Such is logical life. Unfortunately, the Bayesian theory was historically based on the Principle of Indifference: the principle conferred on the theory a supposed status as an *objective* theory of inductive inference. Without it, as we can see in the case of Bayes's derivation, prior probabilities occur in effect as undetermined parameters in the calculation of the posterior distribution. Many Bayesians became preoccupied with the question of whether there is any way of salvaging enough of the substance of the principle to make the enterprise of an objective probabilistic inductive logic worthwhile.

One line of enquiry was conducted by Harold Jeffreys, who investigated which rules, if any, did not yield inconsistencies in the way the Principle of Indifference did under transformations of a continuum-valued hypothesis space like the transformation from p to p' above. Jeffreys was a physicist as well as a probabilist, writing in the first half of the twentieth century, and the problem of finding a rule which can be invariantly (more accurately, *covariantly*[1]) expressed in any co-ordinate system was a familiar one in the physics of general relativity, receiving its mathematical solution in the theory of tensors. Jeffreys found examples which seemed *prima facie* suitable for prior probabilities: the best-known is the rule which can be used wherever there is a statistical model for the observed variate X expressed in a density $f(x \mid \theta_1, \ldots, \theta_k)$. If there is just one unknown parameter θ then this rule assigns as the prior probability $p(\theta)$ the function $\sqrt{I(\theta)}$, or the square root of the *Fisher information*[2], which is

[1] An equation is covariant if it has the same form under all co-ordinate transformation (the terms *covary* in such a way that an identity between them true in one co-ordinate system is true in all; tensor equations famously have this property).

[2] In the multiparameter case the Fisher information is the matrix $(A)_{ij} = \partial^2 L \mid \partial\theta_{ij}$, $i,j = 1, \ldots, k$. Jeffreys's rule assigns the square root of its determinant as the prior probability density.

defined as the expected value, with respect to X, of $-\partial^2/\partial\theta^2$, where L is the logarithm (natural or otherwise; it doesn't matter) of $f(x \mid \theta)$; note that as a result of the differentiation, the expected value may not depend on θ. By elementary calculus the densities $p(\varphi)$ and $p(\theta)$, where is some continuous bijection of θ, are related by the identity

$$p(\theta) = |\, d\varphi/d\theta \,|\, p(\varphi).$$

Hence where the rule for assigning a prior density is given by

$$p(\theta) = \Phi(\theta)$$

for some functional term $\Phi(\theta)$, it is covariant just in case it satisfies the transformation condition

$$\Phi(\theta) = |\, d\varphi/d\theta \,|\, \Phi(\varphi).$$

It is not difficult to show that $\Phi(\varphi) = \sqrt{I(\theta)}$ does indeed transform in just this way, and hence is consistent with respect to arbitrary coordinate transformations.

The Jeffreys rule is far from the only covariant rule for generating prior distributions, though it has received a good deal of attention because some of the priors it generates have been suggested independently. There are technical problems with it, however, principal among which are that (a) the expectation does not always exist, and (b) among the priors that the rule generates are so-called *improper* distributions, that is to say distributions which do not integrate to 1 over their respective ranges ($-\infty$ to $+\infty$ and 0 to ∞) as the probability axioms demand they should. Notable examples of improper distributions are the prior densities generated by it for the mean and standard deviation parameters of a normal model, which are a constant (uniform) and proportional to σ^{-1} (log-uniform) respectively[3]. A good deal has been written

[3] Another feature that some Bayesians take to be problematic is that the rule, depending as it does on the entire outcome space of the experiment, violates the Likelihood Principle (see **5.d** above).

about the status of improper prior distributions, more than we could or should go into here, but some brief comments are in order. Improper distributions might be, and often are, introduced as no more than convenient approximations to proper ones within what is thought to be the likely probable error (the prior proportional to σ^{-1} is approximately that given by the log Cauchy distribution), so long as due care is taken in computations involving them to maintain overall consistency (see, for example, Lee 1997, p.44). That is not the case here, however, where they are deduced *a priori*, and where the fact that they conflict with the standard probability axioms might suggest that the rule is an *ad hoc* device meriting a certain degree of scepticism.

Perhaps this conclusion is premature, since it has been shown that there is a plausible way of amending the axioms which consistently accommodates improper distributions, which is to take a *conditional* probability as primitive. The mathematician Alfred Renyí and the philosopher Karl Popper at about the same time (between 1945 and 1955) independently developed versions of the probability calculus based on a primitive conditional probability function, in which one can in principle have unbounded cumulative distribution functions so long as the conditional probabilities they define remain finite (that is, they can be normalised) (Renyí 1955, p. 295). This development is of considerable interest, without doubt, even though neither Popper's nor Renyí's axiomatisation has been generally adopted (there is perhaps more reason to adopt Renyí's system since it is the closest mathematically to the standard Komogorov axioms, and because Renyí was able to prove the conditions under which his conditional functions are representable as quotients of finite measures (1955, p. 292), which is of course how they are introduced in the standard formalism).

However, even if there are acceptable ways of taming improper distributions, that would still leave entirely open the question why the Jeffreys rule, or any other invariant rule, should be adopted. The choice of any rule would, one would think, need to be justified, and in terms of the logical interpretation we are advocating that means that it has to be justified as a general condition on fair betting odds. The Principle of Indifference looked at first as if it might have a justification of this type, but as it happened it turned

out to be inconsistent. The only justification for Jeffreys's rule that he himself gave was—apart from its demonstrable consistency under transformations of the parameters—its ability to deliver what he considered to be independently desirable priors for certain parameters, in particular—granted the acceptability of improper distributions along the lines mentioned above—for a normal mean and standard deviation separately (but not jointly, since the joint density is proportional to σ^{-2}, which is not the product, as intuitively it should be, of that of the mean by itself and that of the standard deviation by itself) and, to a lesser extent, a binomial probability (Jeffreys 1961, p. 182; p. 184). But in every other respect it seems lacking, a situation not helped by the lack of uniqueness in its invariance property.

9.a.3 Informationlessness

Underlying the attractiveness of the Principle of Indifference is the idea that it allows the data to *speak for itself* by not in any way prejudging the parameter or hypothesis that the data carries information about. Unfortunately, this view is rather obviously untenable. A uniform prior distribution, as we have seen, is uniform across the elements of a particular classification-scheme, for example as determined by the values of a random variable. It therefore says that each element of a basic partition determined by this scheme is equally probable a priori. But any other distribution also gives probabilities of these cells a priori. So *all* prior distributions 'say' something not implied by the observational data alone. In other words, there is *no a priori* distribution that lets the data speak for itself.

Reinforcing this conclusion is the fact, noted in Section 9.a.2, that in continuous spaces we can redescribe the 'elementary possibilities' in terms of another, equivalent, classification scheme in such a way that the original uniform distribution transforms into a highly non-uniform one. Suppose that a quantity x is non-negative, for example. An attitude of prior epistemic neutrality might seem to demand that no one possible value should be regarded as more likely than any other; or in other words, that a uniform distribution is appropriate. But of course there is no proper uniform density on the non-negative line. Worse, a completely neutral attitude would also seem to demand even-handedness with respect to the possible different orders of magnitude of x, those intervals of x determined by the different powers of ten or some other base number. In this case, by the same reasoning, the log-uniform density proportional to x^{-1} is the appropriate one, for this is exactly the density that gives intervals of successive powers equal values. And we have yet another contradiction.

There are no informationless priors, and the quest for them is the quest for the chimera. On the other hand, it might be argued that that quest is arguably little more, if indeed any more, than a quest for *objectivity* in the definition of prior (or indeed any other) probabilities. Whether objectivity really is the unquestionable desideratum it is cracked up to be is an issue we shall come back to later (we shall argue that it is not), but certainly most of the people who have scientific applications of the Bayesian formalism as their methodological goal think that it is. Foremost among these so-called *Objective Bayesians* are (or rather were; one died shortly before the end of the new millennium, the other shortly after) Jeffreys and Jaynes, who explicitly endorsed Jeffreys's general viewpoint and extended some of his methods. Jaynes demanded that prior distributions be objective in the sense that they are "independent of the personality of the user" (1968, p. 117), which he elaborates into the condition that

> *in two problems where we have the same prior information, we should assign the same probabilities.*

But how? Jaynes's answer is to appeal not to the criterion of informationlessness, which is one which, we have seen, cannot be satisfied, but to one of *minimum* information: we should choose the prior containing the *least* information beyond our prior data, or making *the fewest assumptions* beyond that data (1957, p. 623). He argues that this demand can be satisfied in a uniquely determinate way if the background data are of a suitable form, and even where there are no background data. The method that achieves this is, he claims, that of *maximum entropy*. Jaynes draws here on the seminal work of Claude Shannon in information theory. Shannon showed that if $\boldsymbol{p} = (p_1, \ldots, p_n)$ is a probability distribution taking only finitely many values (i.e. it is the distribution of

a random variable X taking n values with those probabilities), then a very plausible set of conditions on a numerical measure of the uncertainty $H(\boldsymbol{p})$ attaching to $\boldsymbol{p}$ determines $H(\boldsymbol{p})$ to be the *entropy* $-\Sigma p_i \log p_i$, unique up to choice of base for the logarithm, which is maximised by the uniform distribution n^{-1}, and minimised by the distribution which attaches the value 1 to a single point. The uncertainty is thus a function of the distribution itself, satisfying the intuitively satisfactory property that the more evenly spread out the distribution, the greater the uncertainty attaching to it.

Given suitable prior information k we should, according to Jaynes, take the prior distribution over X to be that which maximises the entropy subject to the constraints represented by k. Maximising H subject to k means of course that k has to express conditions on $\boldsymbol{p}$ itself. In Jaynes's examples the constraints usually take the form of expectation values 'derived' from very large sets of observations (as arise frequently in experimental physics; other types of constraint might be independence assumptions, conditions on marginal distributions, etc.), and it is well-known that for constraints of this form there is always a unique solution to the maximisation problem (in fact, a unique solution is guaranteed for any set of linear constraints, of which expectations are one type[4]).

It might be objected that outside physics there are few problems amenable to solution this way—where because of the very large numbers involved the data can be represented directly as conditions on the prior $\boldsymbol{p}$. But recall that Bayes's virtual invention of the Bayesian theory as a theory of posterior probabilities was intended to solve the theoretical problem of determining an exact epistemic probability distribution given sample data, no matter how large the sample. Thus the *prior* use of data to furnish constraints on a probability distribution raises the question whether the maximum-entropy method might be in conflict with conditionalisation (given the latter's own conditions of validity; see Chapter 3), a question that others have raised, together with examples where it does seem to be (see Seidenfeld 1979, Shimony

[4] A unique solution is guaranteed for any closed convex set of constraints.

1985), and which seems not to have been satisfactorily answered by advocates of maximum entropy. To say, as Jaynes does, that maximum-entropy is a method for selecting prior distributions, not posterior ones where conditionalisation is the acknowledged tool, rather sidesteps the problem.

That is one question. Another arises in the context of no non-trivial prior information, i.e. no constraints at all other than describing a finite range of values which the data can in principle take (so the constraint set is simply of the form $\Sigma P(x_i) = 1$), the entropy-maximising distribution is the uniform distribution. This might seem—disregarding the problem mentioned above—to amount to a vindication of the Principle of Indifference, since that was postulated on a state of prior ignorance. But it is easy to see that equally it can be regarded as bequeathing to the method of maximum entropy all its most intractable problems, for it implies that we shall get incompatible uniform distributions as maximum-entropy solutions with respect to different ways of representing the space of 'elementary' possibilities (as with the different possibility spaces of L_1 and L_2 above).

Nor can the problem be simply shunted off to one side as a peripheral problem associated with an extreme and arguably rather unrepresentative type of background information. It becomes of central importance when one considers the apparently purely technical problem of extending the Shannon entropy measure to continuous distributions $\boldsymbol{p} = p(x)$. The 'obvious' way of doing this, i.e. by taking H to be the integral

$$-\int p(x) \log p(x) dx,$$

has the serious drawbacks of not being guaranteed to exist and, where it exists, being description-relative: it is not invariant under change of variable (because it is the expected value of a density and hence is dimensional, since it is probability per unit x). The form chosen by virtually all maximum entropy theorists (including Jaynes) to extend H to the continuum, which is invariant under coordinate transformations, is the functional of *two* variables

$$I(\boldsymbol{p},\boldsymbol{q}) = \int p(x) \log [p(x) \mid q(x)] dx,$$

for some other density $q(x)$ (Jaynes 1988, p.124). $I(\boldsymbol{p},\boldsymbol{q})$ is sometimes called the cross-entropy, sometimes the information in $\boldsymbol{p}$ given $\boldsymbol{q}$ (Jaynes simply calls it the entropy). Where the range of X is finite and $p(x)$ and $q(x)$ are discrete distributions with $q(x)$ a uniform distribution $I(\boldsymbol{p},\boldsymbol{q})$ reduces to the ordinary entropy[5]. In addition, since $I(\boldsymbol{p},\boldsymbol{q})$ is never negative, the task is now to minimise a functional rather than, in the case of H, to maximise one.

Adopting I rather than H may solve the mathematical problems, but it leads to a rather obvious interpretative problem: what is $q(x)$? The answer is implicit in the fact that where there is no background information other than that the range of X lies in a bounded interval, the *I-minimising* density is $p(x) = q(x)$ up to a constant factor. In other words, as Jaynes himself pointed out, $q(x)$ virtually has to be interpreted as a density corresponding to (almost) complete ignorance (1968, p. 125), an answer which, as he conceded, appears to lead straight back into the no-go area of the Principle of Indifference. Notice that since $q(x)$ is now necessarily present in every problem where there is a continuous hypothesis-space (as for example there is with most parameter-estimation problems in statistics), the problem of confronting ignorance distributions is no longer confined to a class of cases which can be conveniently written off as merely academic.

It is a testament to Jaynes's ingenuity (and mathematical skill) that he was able to propose a way out of the impasse, and a novel way of formulating the Principle of Indifference which seems to offer at least a partial solution to the problem of transformations. This is his *method of transformation groups*. The idea here is that a definite solution (in $q(x)$) may be revealed by considering equivalent representations of the problem, a class which does not, Jaynes argued, necessarily permit all logico-mathematically equivalent representations of the hypothesis- or parameter-space to be counted as equivalent. More specifically, investigation of the problem may well reveal the only degrees of freedom implicit in its statement are those generated by the action of some transformation group, or class of transformation-groups. To take an example that Jaynes himself uses, suppose that the sampling distribution of an observable variable is known except for a scale parameter σ[6]. We are not completely ignorant about since by assumption we know that it is a scale parameter; but that is all we know. In that case, according to Jaynes, the problem of determining the distribution of σ remains the same problem under arbitrary choices of scale, and these form a group, the group of transformations $\varphi = a\sigma$, $a > 0$. Hence the target distribution $q(\sigma)$ must be invariant under the action of that group. Moreover, since all we know is that σ is a scale parameter, that group determines all the degrees of freedom permitted by the problem: it is not permissible to demand invariance under taking logarithms, for example.

The following simple argument now shows that q is uniquely determined by this condition. Firstly, by the probability calculus and elementary differential calculus the prior densities $q(\sigma)$ and $h(\varphi)$ are related by the equation $(d\varphi / d\sigma)h(\varphi) = q(\sigma)$, *by*

$$ah(a\sigma) = q(\sigma).$$

The assumption of scale-invariance means that the scale-shift from to leaves the form of the prior unchanged: h and q are exactly the same function on their common domain, the set of positive real numbers. Substituting $h(\,.\,) = q(\,.\,)$ accordingly, we obtain

$$aq(a\sigma) = q(\sigma)$$

Setting $\sigma = 1$ we have $aq(a) = q(1)$ for all $a > 0$, and hence we infer that $q(\sigma) \propto \sigma^{-1}$. In other words, we have, up to proportion-

[5] There are several derivations from first principles of $I(\boldsymbol{p},\boldsymbol{q})$ as a measure of relative information whether $\boldsymbol{p}$ and $\boldsymbol{q}$ are discrete or continuous.

[6] A parameter of a sampling distribution is called a scale parameter if the data transformation taking y to $x = cy$ for a constant c has the effect of changing the likelihood function so that the likelihood of λ on y is the same as that of $c\lambda$ on x (to avoid the problem that the likelihood function is defined only up to a multiplicative constant, the function referred to here is the so-called standardised, i.e. normalised, likelihood, which is any version $L(\lambda \mid y)$ divided by $L(\lambda \mid y)d\lambda$). λ is a *location parameter* if the transformation $x = y + c$ gives $\lambda + c$ the likelihood on x corresponding to that of on y. Familiar examples of each type of parameter are the standard deviation of a normal distribution (scale parameter), and the mean (location parameter).

ality, determined the form of the (improper) prior density to be log-uniform (in fact, this density is invariant under a much larger class of transformations; namely, all those of the form ax^b, $a > 0$, $b \neq 0$).[7]

This use of the statement of the problem to determine the precise extent to which priors should reflect ignorance suggested to Jaynes that in other problems where the Principle of Indifference merely yields inconsistencies his own method might be successful. The most celebrated of these was Bertrand's celebrated, or rather notorious, inscribed triangle problem. This is the problem of determining the probability that a 'randomly' drawn chord to a circle will have a length less than the length L of the side of the inscribed equilateral triangle (of course, there is an uncountable infinity of inscribed triangles, but they are all congruent). The paradox arises because there are different, but equivalent, ways of specifying a chord of a given length. Here are three Bertrand considered (the exposition follows Gillies 2000, pp. 37–39): (1) where the midpoint of the chord occurs on a radius (S_1); (2) the angle between an end-point of the chord and the tangent at that point (S_2); and (3) the position, in polar coordinates say, of the midpoint in the circle (S_3). S_1, S_2 and S_3 determine different possibility-spaces: (1) all the points on the radius; (2) the angles from 0 to π, and (3) the set $\{(r,\theta)\colon 0 \leq r \leq R,\ 0 \leq 2\pi\}$, where R is the length of the radius. Subsets T_1, T_2 and T_3 of S_1, S_2 and S_3 respectively correspond to the event, call it C, that the chord-length is less than L. T_1 is the set of points on a radius between its perpendicular intersection with a side of the inscribed triangle and its intersection with the circumference of the circle; T_2 is the 60° angle between the tangent and the nearest side of the inscribed triangle whose vertex is at the tangent: and T_3 is the set of points in the interior circle of radius $R/2$. Both members of the pairs (S_i,T_i) have a 'natural' measure m_i induced by their geometrical structure: in the first, length, in the second, angular measure, and in the third, area

[7] Scale-invariance has also been used to explain why the leading digit n in the measurements in significant figures of many naturally occurring magnitudes often has the logarithmic distribution $\log((1 + n) / n)$ (a regularity known as *Benford's Law*; see Lee 1997, pp. 100–02).

measure. The Principle of Indifference decrees that relative to each such pair the probability of C is equal to $m_i(T_i) \mid m_i(S_i)$. Thus, the probability of C relative to the first pair (S_1,T_1) *is* 1/2, relative to the second is 1/3, and relative to the third is 1/4. Thus we have *Bertrand's Paradox*: relative to three different, apparently equivalent, ways of describing the problem, we get three different values.

According to Jaynes, however, Bertrand's problem does admit a unique solution, $\frac{1}{2}$. He proposes that we view the problem in the context of somebody, say Jaynes, throwing longish straws randomly so that they intersect the circumference of a circle inscribed on a plane surface. Suppose $f(x,y)$ is the probability-density that the midpoint of the chord lies at the point (x,y) in or on the circle (assumed to be centred at the origin). Nothing in the problem specifies the origin, the orientation of the coordinates, or the size of the circle, so according to Jaynes this tells us that f should be translation, rotation and scale-invariant. Rotational invariance implies that f depends on (x,y) through $\sqrt{(x^2+y^2)}$ only. Transforming to polar coordinates, the joint probability density $p(r,\theta)$ must therefore be $rf(r)$, $0 \leq r \leq R$, $0 \leq \theta \leq 2\pi$. Now consider a scale transformation $r' = ar$, $a \leq 0$, which we can think of as blowing up or shrinking the original circle. Suppose it is shrunk, i.e. $a \leq 1$. The probability densities $p(r)$ and $p'(r')$ are related in the usual way by

$$p'(r')\, dr'/dr = p(r) \qquad \text{(i)}$$

We cannot immediately invoke scale invariance to identify the two density functions p and p', since they have different domains, $[0,aR]$ and $[0,R]$ respectively. What we can do, though, is to restrict r to $[0,aR]$ by conditioning p on the information that r lies in $[0,aR]$, and then equate the resulting *conditional* density with p'. This is what Jaynes does, and differentiates the resulting equation with respect to a to obtain a differential equation whose solution is

$$f(r) = qr^q \quad {}^2/2\pi R^q$$

where $0 \leq q$. Finally, Jaynes shows that translation-invariance uniquely determines $q = 1$, giving $f(r) = 1/2\pi Rr$ (however, as he

points out, translation-invariance is by itself so strong a condition that it already determines the result). Thus the required probability density is uniform, and the probability that the chord is less

than the length of the inscribed triangle is $1 - (1/2\pi R)\int_0^{R}\int_0^{2\pi} dr d\theta$

which is easily seen to be equal to $\frac{1}{2}$, i.e. to the first of the solutions above. Not only, apparently, is the solution uniquely determined, but it is one which, as Jaynes noted with satisfaction when he carried out experiments of actually throwing straws into a circle, was observed to "an embarrassingly low value of chi-squared" (1973, p. 487).

Jaynes conceded that not all the examples where the Principle of Indifference breaks down yield to this sort of attack: he cites the well-known wine/water problem (von Mises 1939, p. 77) as one which does not. But even his solution of the Bertrand problem raises serious questions. According to Jaynes, the method of transformation groups consists of noting those aspects of the problem which are left unspecified, and then requiring invariance of the prior under the corresponding groups of transformations (1968, p.128, 1973, p.430). But note that in Jaynes's own statement of Bertrand's chord problem, which actually specifies an individual, J, throwing straws into a circle inscribed on a plane surface, there is no mention of the relative speed of J and the circle. Suppose that they are in uniform motion with velocity v, and that the coordinate frames of J and the circle are coincident when J notes the midpoint of the chord. It follows from the equations of the Lorentz transformation that the circle in the frame taken to be at rest (say, J's) is an ellipse in the moving frame, with eccentricity v (in units of $c = 1$), and minor axis in the direction of motion. Thus there is no longer rotational symmetry for J, since r is shortened in J's frame by the factor $[(1 - v^2)\cos^2\theta + \sin^2\theta]^{1/2}$. The problem seems no longer so well-posed, at any rate for J.

Jaynes would no doubt reply that it was intended that the circle be at rest (at any rate up to what is observationally detectable) with respect to J. But according to Jaynes's own criterion, changing any characteristic not actually specified in the statement of the problem should count as giving an equivalent problem, and the relative speed of J and circle was not mentioned in his reformulation of Bertrand's problem. Even if it were fixed by some additional stipulation, that would still never succeed in reducing all the degrees of freedom to a manageable number, since they are in principle infinite. Why shouldn't arbitrary variations in space-time curvature be included in the list, or even arbitrary coordinate transformations, which are not mentioned either? Of course, one could simply state the 'permitted' transformations at the outset, but even if they generate a unique solution it will be tantamount to invoking the sort of personal decision that the idea of objective priors was intended to replace:

> if the methods are to have any relevance to science, the prior distribution must be completely "objective", in the sense that it is independent of the personality of the user ... The measure of success ... is just the extent to which we are able to eliminate all personalistic elements and create a completely "impersonalistic" theory. (1968, p. 117)

Jaynes's entropy-based methodology was an attempt to bypass the problems facing the Principle of Indifference: instead, the problem of determining the distribution $q(x)$ merely meets them all again head-on. They are not to be by-passed so easily. Recall Jaynes's claim that, because it maximises uncertainty, the maximum entropy distribution, where it exists, makes the fewest commitments or assumptions beyond those in the data. Even granted the equation of entropy with uncertainty this assertion rather obviously begs the question, and is certainly incapable of proof, or even moderately convincing supporting argument, because on any reasonable understanding of the terms involved it is false. A flat distribution maximises uncertainty, but it commits its user to certain views about the probability of every outcome, or neighbourhood of every outcome if it is continuous. As far as its probabilistic content is concerned, therefore, it makes *exactly* as many commitments as any other distribution. This conclusion is, of course, just a reprise of our verdict on the search for informationless priors (there are none, because every distribution contains just as much probabilistic information as any other). Unfortunately, it has yet to be drawn by some authoritative Bayesians:

> Even when a scientist holds strong prior beliefs about the value of a parameter, nevertheless, in reporting his results it would usually be appropriate and most convincing to his colleagues if he analyzed the data against a *reference* prior [a reference prior is a prior distribution which is slowly varying and dominated by the likelihood function in the region where the latter takes its largest values]. He could then say that, irrespective of what he or anyone else believed to begin with, the prior distribution represented what someone who *a priori* knew very little about should believe in the light of the data. (Box and Tiao 1973, p. 22.)

There do admittedly exist alleged *proofs* that the maximum entropy method is uniquely determined by plausible assumptions; indeed, there is "quite an industry", to quote Paris (1994, p. 79). But closer investigation reveals that there is often (much) more to these conditions than meets the eye. For example, Paris and Vencovská's (2001) apparently innocent 'renaming principle' implies that, with a constraint set asserting that with probability one there are n elementary possibilities, each should receive probability $1/n$ (and we are back with the Principle of Indifference in yet another guise); another of their conditions is that if the constraint contains only conditional probabilities of a given c and b given c, and the unconditional probability of c, then a and b should be independent given c, a very strong condition reminiscent of the default-negation rule in logic programming.[8]

There is one other aspect of maximum entropy that we should mention before leaving the subject, though this concerns exclusively the Minimum Information Principle under the very different interpretation *as an updating rule*, from Jaynes's. On this interpretation, $q(x)$ is the prior probability and $p(x)$ the posterior updated on the constraints relative to which I is minimised. This conveniently sidesteps the problems with Jaynes's use of maximum entropy/minimum I as a determiner of prior distribitions. With the I-minimising $p(x)$ treated as a new way of generating a

[8] A theorem of Williamson, about separating sets of variables in a constraint graph, implies that the maximum entropy solution for the constraints above automatically renders a and b conditionally independent given c (2005, p. 86). This underlines how far the maximum entropy method transforms null information into positive information.

posterior probability relative to $q(x)$'s prior, and the minimum-information principle thereby transformed into an updating rule, things look altogether more promising. For a start, the Principle subsumes both Bayesian and Jeffrey conditionalisation: minimising $I(\boldsymbol{p},\boldsymbol{q})$ subject to a shift from $q(e)$ to $p(e) = p$ yields the latter:

$$p(x) = q(x \mid e)p + q(x \mid -e)(1 - p)$$

(we get the appropriate general form when the shift occurs simultaneously on the members of an arbitrary finite partition; for a full discussion see Williams 1980). Where the contraint is that $q(e)=1$, we obtain Bayesian conditionalisation: $p(x) = q(x \mid e)$.

These facts provide another defence of conditionalisation, Bayesian or Jeffrey, depending on the constraints. If we grant that $I(\boldsymbol{p},\boldsymbol{q})$ can be interpreted as a distance measure in distribution space between $\boldsymbol{p}$ and $\boldsymbol{q}$, then minimising $I(\boldsymbol{p},\boldsymbol{q})$ for fixed $\boldsymbol{p}$ means, on this view, selecting as one's new distribution that which is as close to $\boldsymbol{p}$ as possible subject to the constraints. Thus conditionalisation becomes justified as the selection of the distribution closest to the prior. There are several problems with this defence, however. One is that $I(\boldsymbol{p},\boldsymbol{q})$ is no ordinary distance measure since it is not symmetric. Another is that there are alternative metrics which do not endorse conditionalisation in this way. A deeper objection is that it simply begs the question *why* one should choose the closest measure to $\boldsymbol{p}$, particularly as the shift on e might be very considerable. Shore and Johnson (1980) prove that the information-minimising measure is the only one satisfying a list of what they term consistency constraints, but one of these is the assumption that the task is to extremise a functional (why should this be a consistency condition?), while another turns out to be a strong independence condition (and rather similar to the independence condition of Paris and Vencovská discussed in the preceding paragraph) which is certainly not a mere consistency constraint and, as Uffink points out with supporting examples, may in the appropriate circumstances be very unreasonable (1995, pp. 245–247; Uffink's article also contains an excellent critical discussion of Jaynes's use of Maximum Entropy). Given the other problematic features of conditionalisation we pointed to in Chapter 3, we feel that in linking its fortunes to the principle of

minimum information no real advance has been made in justifying its adoption as an independent Bayesian principle.

9.a.4 Simplicity

No examination of ways of trying to impose 'objective' constraints on prior probabilities is complete without a discussion of a criterion of ancient pedigree: *simplicity*. Relativised to the Bayesian way of looking at things, the idea is that greater relative simplicity of an explanatory or predictive theory should be reflected in its having a higher prior probability. But the criterion, while plausible, has its problems. One is lack of univocality: there are different ways in which we judge things simple, and these are not all equivalent, and some are highly description-relative. For example, the equation of a circle with radius k centred at the origin has the equation $x^2 + y^2 = k^2$ in Cartesian co-ordinates, but the apparently much simpler equation $r = k$ in polar coordinates. In general there are many different ways of characterising the main principles of a theory, whose choice may depend on a variety of factors, and which may seem more or less simple depending on the application to hand.

But there is a more objective, less language-dependent sense of simplicity, which also appears to play a role in at least some areas of science, and that is simplicity in an Occam's razor-sense (Occam's famous methodological principle was 'entities are not to be multiplied without necessity'), which has a precise mathematical expression as *fewness of independent adjustable parameters*. This certainly strikes a respondent chord with scientists: one of the considerations telling against the current Standard Model of subatomic physics is that it contains no fewer than 20 adjusted parameters. Jeffreys's modified Galileo law (4.i above), which no-one would accept in preference to Galileo's own, has the form of Galileo's law with k additional adjustable parameters evaluated at the data points $t_1, \ldots, t_k$. A simpler form which fits the data as well is not only more elegant and mathematically tractable but also, we feel, *more likely on that account to be true*. In that case, why not adopt as an independent rule that hypotheses with fewer adjustable parameters should receive greater prior probability? This is what Jeffreys himself advocated, calling the rule the

Simplicity Postulate (1961, pp. 46–50). It may not determine prior probabilities uniquely, but it does act as an objective constraint on them where it is applicable.

Rather surprisingly, some have argued that such a principle is actually inconsistent. Popper was probably the first to make this claim (1959a, pp. 383–84), and Forster and Sober in a series of papers in effect repeat Popper's argument. This is that since a polynomial relation of degree n is also one of every higher degree $m>n$, with the coefficients of all terms of degree greater than n set equal to zero, the lower-degree hypothesis cannot have a larger probability than any of higher degree, since the first entails the second and probability must respect entailment (Sober and Forster 1994). For example, a straight line $y = mx + c$ is also a parabola with the coefficient of x^2 set equal to 0.

The argument is, however, easily rebutted by noting that the interest is usually in testing against each other not compatible but *incompatible* hypotheses, for example whether the data are better explained by the existing hypothesis or by adding a new parameter in the form of a *nonzero* coefficient to a higher-degree term (Howson 1988a). Thus, to use Forster and Sober's notation, suppose LIN is the set of all linear models and QUAD the set of all quadratic ones. In testing whether the true model is a linear one M_L or a quadratic one M_Q the tester is *not* testing LIN against QUAD; since they have common elements it would be like testing M_L against M_L. The test is between LIN and QUAD* where QUAD* contains all the models in QUAD which are not in LIN. While P(LIN) is necessarily no greater than P(QUAD) by the probability calculus, P(LIN) can consistently be greater than P(QUAD*).

Jeffreys himself regarded discriminating between such disjoint families as LIN and QUAD* in curve-fitting problems as a classic arena for the Simplicity Postulate, pointing out that the penalty of being able to fit the data exactly by means of a plentiful enough supply of free parameters is *overfitting*:

> If we admitted the full n [parameters] . . . we should change our law with every observation. Thus the principle that laws have some validity beyond the original data would be abandoned (1961, p. 245)

Indeed, "the simplest law is chosen because it is the most likely to give correct predictions" (Jeffreys 1961, p.4). Since the promotion

of predictive accuracy is Forster and Sober's own declared aim, their charge that Jeffreys's restriction of the simplicity ordering to disjoint polynomial families is an "ad hoc maneuver" which "*merely changes the subject*" (1994, p. 23; their italics) is simply incorrect.

Forster and Sober make a further curious claim. The hypothesis that some relationship is a particular degree of polynomial asserts that it is some (unspecified) member of the corresponding family of curves, and hence computing its posterior probability means computing the posterior probability, and hence the likelihood, of that family. On this point Forster and Sober claim that

> it remains to be seen how [Bayesians] . . . are able to make sense of the idea that families of curves (as opposed to single curves) possess well-defined likelihoods. (1994, p. 23)

This is strange, because the ability to make sense of the idea is *guaranteed* in the Bayesian theory, whereas, ironically, Forster and Sober's charge is rather accurately brought against their own account: in any theory restricted to considering likelihoods there is no comparable body of first principles which generates likelihoods of families (disjunctions) of hypotheses. Even 'the father of likelihood', R.A. Fisher, conceded that it makes no sense to talk of the likelihood of a disjunction (he described it as like talking about the "stature of Jackson or Johnson" (1930, p. 532). In the Bayesian theory the likelihood of a disjoint family $\{h_i\}$ of curves (determined, say, by a discrete parameter) with respect to data e is easily obtained via the probability calculus in terms of prior probabilities and the likelihoods of each member of the family

$$L(\{h_i\}|e) \propto P(e|vh_i) = \frac{\Sigma P(e \mid h_i)P(h_i)}{\Sigma(h_i)}$$ (where vh_i is the disjunction of the h_i. In the continuous case the sum is replaced by integration where the integrals are defined. Consider, to take a simple example, the hypothesis that the data $\boldsymbol{x}$ are normally distributed with standard deviation around some unknown value t. The likelihood is equal to $\int\sqrt{(2\pi\sigma^2)^{-1}}\exp[-(1/2\sigma^2)(\boldsymbol{x}-t)^2]f(t)dt$, where $f(t)$ is the prior density and the integration is over the range

of values of t. The prior may of course be subjective, but that still does not make the likelihood *ill-defined*; on the contrary, formally speaking it is perfectly well defined.

While, *pace* Forster and Sober, there is no *technical* problem with a Simplicity Postulate, we doubt that simplicity, in itself and divorced from the consideration of how it is related to the scientist's background information, is or should be regarded as a criterion of any great importance in guiding theory-choice. While a dislike of too many adjustable parameters is often manifested by scientists, it arguably depends on the merits of a particular case, and a closer examination usually reveals that it is considerations of *plausibility*, not simplicity in itself, that ultimately determine attitudes. Indeed, it is easy to think of circumstances where the simplest hypothesis consistent with the data would be likely to be rejected in favour of a more complex one: for example, even before the data are investigated in some piece of economic forecasting, models with very few parameters would be regarded with great suspicion precisely because there is an equally strong belief in a (large) multiplicity of independent causes.

This is not to say that nothing definite can be said in favour of a here-and-now preference for theories with fewer parameters. For example, it can easily be shown that, other things being equal, hypotheses with fewer parameters get better confirmed by predicted observational data: a certain amount of the data is absorbed merely in evaluating the parameters, leaving the remainder to do the supporting—or not—of the resulting determinate hypothesis (Howson 1988b, pp. 388–89). But 'other things being equal' here means that the two hypotheses have roughly equal prior probabilities. This is an important point in the context of the perennial 'accommodation versus prediction' debate, since what it points to are circumstances where the accommodating hypothesis is more highly confirmed than the independently predicting one, namely where the prior probability of the latter is sufficiently low compared with that of the former (this is shown in detail in Howson, op. cit.). Everything, in other words, over and above fit with the data is a matter of prior judgments of plausibility.

And this seems true even at a very intuitive level. A curve of high degree in some 'natural' co-ordinate system would be thought by most people to be more complex than one of low

degree. But suppose a highly complex relationship $y(x)$ is thought to hold between two observable variables x and y. Draw an arbitrary straight line through the curve intersecting it at k points (x_i,y_i), $i = 1, \ldots, k$. Now empirically determine the values $\boldsymbol{y}(x_i)$ for the selected x_i. Suppose that, within the bounds of error, $y_i = \boldsymbol{y}(x_i)$. Would you regard the straight line as being confirmed by these data rather than the favoured hypothesis? No. Ultimately, the criterion counting above all is plausibility in the light of background knowledge; to the extent that simplicity is a criterion, it is to the extent that it supervenes on prior plausibility, not the other way round.

There does however remain the legitimate methodological concern that too plentiful a use of adjustable parameters to fit current data ever more exactly invites the risk of future overfitting, "as a model that fits the data too closely is likely to be tracking random errors in the data" (Myrvold and Harper 2002, p. 137). But then again, a very poor fit to present data with n parameters suggests that it may be necessary to introduce an $(n + 1)$th. It is all a question of balance. It might be thought that this rather modest conclusion sums up all that can usefully be said on the matter, but Forster and Sober, scourges of Bayesianism, claim that a theorem of Akaike shows that there is actually a great deal more that can be said, and of a mathematically precise character. To examine this claim we need first a precise notion of *fit to the data* of a hypothesis with m free parameters determining (it is the 'disjunction' of all of them) a corresponding family h_m of specific hypotheses, which we shall take to be characterised by the density $p(y \mid q)$ with q an adjustable parameter-vector of dimensionality m. For any given data x let $\max_x(m)$ be that member of h_m which has maximum likelihood on x, i.e. the hypothesis whose parameters are the maximum-likelihood estimates $\hat{\theta}(x)$ determined by the data x. Let us, for the sake of argument, follow Forster and Sober and regard this as the hypothesis in h_m which best fits x. Now let $\max_x(m)$ be the log-likelihood of xm on x. Note that formally $\max_x(m)$ is a function of x. Let P^* be the true probability distribution of the possible data generated by the observations. Forster and Sober take the joint expectation $E^*_y E^*_z[\log p(y|\theta(x)]$ computed relative to P^* to measure the predictive accuracy of h_m. Again, for the sake of argument let us agree, though we shall have something to say shortly about this use of expectations.

The next step is the dramatic one. According to Sober and Foster, Akaike has shown how to construct *a numerically precise estimate of the degree to which additional parameters will overfit*. For Akaike showed that under suitable regularity conditions $l(\max_x(m) - m$ for the actual data $\boldsymbol{x}$ is an unbiased estimate of the predictive accuracy of h_m, as defined above. Akaike's result, often known as AIC for 'Akaike Information Criterion', thus appears to tell us that in estimating the predictive accuracy of a hypothesis with m free parameters from the current data, we must subtract from its current goodness of fit a penalty equal to m. In other words, other things being equal (current fit), we do better from the point of view of future predictive success to choose the simpler hypothesis.

What should we make of this? One thing we are not questioning is the connection with information theory. Akaike also showed that AIC is an estimate of the discrepancy between probability distributions as measured by the Kullback-Leibler measure of distance between a model and the 'true' distribution, a measure which we have already encountered in the discussion earlier of the Principle of Minimum Information (hence 'Akaike Information Criterion'). This is eminently discussible, in particular because (i) the Kullback-Leibler measure is not actually a true distance (it is not symmetric), and (ii) there are other discrepancy measures in distribution space, for example variation distance, but the Akaike criterion fails to estimate these, that is, it is not robust over discrepancy measures. But that is not our present concern. That concern is that Sober and Forster see AIC as justifying the claim that on average simpler models are predictively more accurate than complex ones. We shall now say why we think there are serious grounds for questioning this claim.

Firstly, there is the question of how well the 'suitable regularity conditions' are satisfied in typical problems of scientific theory choice; the conditions are actually rather restrictive (see Kieseppä 1997 for a fuller discussion). Secondly, this account suffers from an acute version of the reference-class problem. Suppose I use as many parameters as is necessary to obtain a perfect fit to the data within some polynomial family. The resulting hypothesis h is a member of the singleton family $\{h\}$, which is a family with no adjustable parameters and which has excellent fit

to the current data. According to the Akaike criterion no other family can do better than $\{h\}$. But this is absurd, because we know that h will overfit, and indeed is exactly the type of hypothesis whose merited killing-off this analysis was developed to justify.

There is an extensive literature on Forster and Sober's advocacy of the Akaike criterion, and they themselves do attempt to answer this rather devastating objection. We do not think they succeed, but there is another type of objection to their enterprise which we believe to be just as undermining. An unbiased estimate, recall, is characterised in terms of its expected value. The use of estimates based on their expected values is usually justified in terms of an asymptotic property: the sample average converges probabilistically to the expectation under suitable conditions. This is the content of a famous theorem of mathematical probability known as Chebychev's Inequality, and Bernoulli's theorem is a famous special case of it, where the expected value of the sample average (relative frequency) is the binomial probability. But as we saw, you cannot straightforwardly infer that probability from an observed relative frequency; indeed, the only way you infer anything from the relative frequency is via the machinery of Bayesian inference, and there is none of that here. Moreover, it is easily shown that there is an *uncountable infinity* of unbiased estimates of any given quantity, all differing from each other on the given data. How can they all be reliable when they contradict each other? And to compound the problem still further is the fact that the quantity being estimated by the Akaike criterion is yet another expectation!

Frequentists usually respond to the question raised by the multiplicity of unbiased estimators by saying that one should of course choose that with minimum variance; if the variance is very small then—so the argument goes—one is more justified in using that estimator. But even were one to grant that, there is no guarantee that the Akaike estimator is minimum variance, let alone small. Even if it were small, that would still leave the question of how likely it is that this value is the true one, for which, of course, one needs the Bayesian apparatus of posterior probabilities.

Perhaps surprisingly, there is a Bayesian analogue of Akaike's criterion, due to Schwarz and known as BIC (Bayesian Information Criterion). BIC replaces the penalty m in Akaike's estimator by the quantity $(1/2)m\log n$, where n is the number of independent observations, but is otherwise the same:

$$l(\max_x(m)) - (1/2)m\log n \qquad \text{BIC}$$

But the justification of BIC, as might be expected, is very different from that of AIC: BIC selects the model which maximises posterior probability as the data grows large without bound (Schwarz 1978, p. 462). Under fairly general conditions the posterior probability takes its character from the behaviour of the likelihood function for a sufficiently large independent sample, a fact which explains the presence of the likelihood term in BIC.

That sounds good, but unfortunately BIC has its own attendant problems, principal among which is that its justification is asymptotic, giving the model with the highest posterior probability but only in the limit as the data extends. But we do not live at asymptotes. It is a simple matter to choose other sequences depending on the data which have the same limit properties, but which are all very different on any finite sample. Again, there is nothing like a guarantee from BIC that using it at any given point in the accumulation of data we are on the right track. From a purely theoretical point of view, however, BIC does, in a way that AIC does not, offer an intelligible and straightforward justification of the intuition that too-precise fitting to the data means overfitting which means that it is unlikely that the result will survive a long enough run of tests. Thus Jeffreys's claim that 'the simplest law . . . is the most likely to give correct predictions' is clearly underwritten by BIC, at any rate asymptotically. AIC, by contrast, merely tells us that expected fit is improved, according to an unbiased estimator, by reducing the number of free parameters, and as we have seen it in fact tells us nothing about how much more likely a simple theory is to be true.

And now we are back on familiar ground, where the choice between two accounts of why simple hypotheses are meritorious is at bottom just the choice between a frequentist, classical view of statistical inference as against a Bayesian. It has been the burden of this book that only the Bayesian offers a coherent theory of valid inductive inference, and that, despite its suggestive terminology, of unbiasedness, sufficiency, consistency, significance and the like, the classical theory is in fact shot through with sys-

tematic question-begging. The discussion of these two superficially similar but in reality very different justifications of simplicity underlines that view.

AIC and BIC are actually not the only approaches to trying to justify simplicity-considerations in terms of some more tangible methodological goal, though these deal with rather different conceptions of simplicity. One, due to Solomonoff and others, appeals to Kolmogorov complexity theory. Assume that hypotheses assign probabilities to possible data strings. We can suppose without loss of generality that both hypotheses and data are coded as finite strings of 0s and 1s. The complexity of any such string is defined to be the length of the shortest computer program which will generate it. The fact that such program-lengths across different 'universal' programming languages (like LISP, PROLOG, JAVA etc.) can be proved to be uniformly bounded by a constant means that the definition is to that extent relatively language-independent. Suppose a data sequence of length n is observed, and that P is the true distribution of the data. Let the error in predicting the next member of the sequence between P and any other hypothesised distribution, P', be the square of the difference between the two probabilities on that member conditional on the data. Solomonov showed under quite general conditions that if a certain prior distribution λ ('the universal enumerable semi-measure') is employed, which is also a prior distribution weighting complex sequences lower than simpler ones according to the complexity criterion above, then the expected value of the error converges to zero (Li and Vitányi 1997). But now we have yet another criterion justified in terms of an expected value, and everything we said earlier about the Akaike criterion applies here equally.[9]

9.b | Summary

Simplicity is a snare, in our opinion, in whatever formal guise. Ultimately, it is *plausibility* that is the issue, and this does not always harmonise with what the a priori theorist takes as his or her ideal of simplicity. The same general observation also, in our opinion, undercuts the quest for objective priors in general. Our view, which we have stated several times already (but see that as no reason to stop!) and believe the most natural way of interpreting the Bayesian formalism, is that the latter is simply a set of valid rules for deriving probabilistic consequences from probabilistic premises. If you want accurate conclusions you should make your assumptions as accurate (in your own eyes) as you can. But the objectivity of the enterprise consists in the objective validity with which you draw conclusions from those assumptions.

In this view the quest by many for 'objective' prior distributions is not only unnecessary but misconceived, a conclusion is reinforced by the problems which arise in pursuing that quest, and which seem to be resolvable only by the sorts of ultimately subjective decision that makes the enterprise self-defeating. People, even those possessing the same background information, and even experts, may still have different opinions, *pace* Jaynes. Trying to force this, in our view entirely legitimate, diversity of opinions into a single uniform one is misguided Procrusteanism, and would have deleterious consequences for the progress of science were it to be legislated for.

It is, in addition, certainly not sensible to throw away relevant information, yet this is in effect just what is recommended by those who tell us that we should always use reference priors wherever possible, or give the simplest hypotheses the highest a priori probability. But none of this means that the Bayesian theory without 'objective' priors effectively imposes no constraints at all (as has often been charged). On the contrary, the consistency constraints represented by the probability axioms are both stringent and very objective, as stringent and objective as those of deductive logic. And in a theory of valid inference that is not only as good as it gets, but quite good enough.

9.c | The Old-Evidence Problem

Or is it good enough? There will always be people who object to any theory, and the Bayesian theory is no exception to this rule.

[9] In fact, model selection criteria of the sort we have mentioned (and others) are extensively used (see, for example, Burnham and Anderson 2002).

We have tried to deal with the objections which we feel merit serious discussion. We shall end, however, with one that doesn't, but we shall take a look at it nonetheless because it is often seriously advanced as the most serious objection to using the theory of personal probability in any methodological role.

It goes as follows. The Bayesian theory is supposed to reflect patterns of accepted reasoning from data in terms of the way the data change one's probabilities. One type of such reasoning is assessing the impact of evidence on a hypothesis of data obtained before the hypothesis was first proposed. The stock example is the anomalous precession of Mercury's perihelion, discovered halfway through the nineteenth century and widely regarded as supporting Einstein's General Theory of Relativity (GTR) which was discovered (by Einstein himself) to predict it in 1915. Indeed, this prediction arguably did more to establish that theory and displace the classical theory of gravitation than either of its other two dramatic contemporary predictions, namely the bending of light close to the sun and the gravitational red-shift. But according to nearly all commentators, starting with Glymour 1980, this is something which *in principle* the Bayesian theory cannot account for, since e is known then $P(e) = 1$ and it is a simple inference from the probability calculus that $P(h \mid e) = P(h)$; i.e., *such evidence cannot be a ground for changing one's belief in h.*

Despite all this, the 'old evidence' objection' is not in fact a serious problem for the Bayesian theory; indeed, it is not a problem at all, certainly in principle. What it really demonstrates is a failure to apply the Bayesian formulas sensibly, and to that extent the 'problem' is rather analogous to inferring that $3/2 = x/x = 1$ from the fact that $3x = 2x$ if $x = 0$. To see clearly why we need only note an elementary fact about evidence, which is that *data do not constitute evidence for or against a hypothesis in isolation from a body of ambient information.* To talk about e being evidence relevant to h obviously requires a background of fact and information against which e is judged to be evidence. A large dictionary found in the street is not in itself evidence either for or against the hypothesis that Smith killed Jones. Relative to the background information that Jones was killed with a heavy object, that the dictionary belonged to Smith, and that blood found on the dictionary matches Jones's, it is. In other words,

'being evidence for' connotes a *three-place relation*, between the data, the hypothesis in question, *and* a body k of background information. The evidential weight of e in relation to h is assessed by how much e changes the credibility of h, in a positive or negative direction, *given* k.

Clearly, a condition of applying these obvious criteria is that k does not contain e. Otherwise, as the old-evidence 'problem' reminds us, e could not *in principle* change the credibility of h: requiring that k not contain e, before judging its evidential import relative to k is merely like requiring that the car engine is not already running in any test to see whether a starter motor is working properly. Granted that, we can see that the old-evidence 'problem' really is not a problem, merely an implicit reminder that if e is in k then it should first be deleted, as far as that can be done, before assessing its evidential weight.[1]

It is often objected against this that there is no uniform method for deleting an item of information from a database k, and often it seems that there is no way at all which does not represent a fairly arbitrary decision. For example, the logical content of the set $\{a,b\}$ is identical to that of $\{a, a \rightarrow b\}$, where a and b are contingent propositions, but simply subtracting a from each will leave two different sets of consequences; b will be in the first and not the second, for example, if the sets are consistent. Much has been made of this problem, and some have been led to believe that the task is hopeless. Fortunately, this is far from the truth. Suzuki (2005) has shown that there are consistent probabilistic contraction functions which represent the deletion of e from k relative to plausible boundary conditions on such functions (these conditions are furnished by the well-known AGM (Alchourrón-Gärdenfors-Makinson) theory of belief-revision; see Suzuki op. cit. for references). The exhibition of a *particular*

[1] Given the routine dismissal of the counterfactual move in the literature, readers may be surprised to learn that it is in fact standard Bayesian procedure. Once any piece of evidence is 'learned' it becomes 'old', and according to those who advance the 'old evidence problem' as an objection to the Bayesian methodology, it should no longer confirm any hypothesis (indeed, conditionalising on e automatically takes its probability to 1). So to regard any evidence, once known, as confirming one has to go counterfactual.

probability function representing the deletion of e will in general reflect the way the agent her/himself views the problem, and it is completely in line with the personalistic Bayesian theory adopted in this book that the request for an objective account of how this should be done is simply misplaced. Nevertheless, it can often be expected that the constraints imposed by the background information will practically determine the result, and this is certainly true for the example which prompted the discussion, the observation of the precession of Mercury's perihelion, as we shall now show.

The discussion follows Howson 2000, p.194. We start, appropriately, with Bayes's Theorem, in the form:

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

where as before h is GTR, e is the observed data on Mercury's perihelion (including the error bounds), $p = P(h)$, and P is like the agent's probability function except that it does not 'know' e. Following category theory, we could call P the 'forgetful functor', meaning in this case that it has 'forgotten' e. We shall now show that, despite this idea sounding too vague to be useful, or even, possibly, consistent, *the data of the problem are sufficient to determine all the terms in the equation above, at any rate to within fairly tight bounds*.

Firstly, we have by assumption that h, together with the residual background information which P is assumed to 'know', entails e, so $P(e \mid h) = 1$ by the probability axioms *independently of any particular characteristic of P*. Thus the equation above becomes

$$P(h \mid e) = \frac{p}{p + P(e \mid \sim h)(1 - p)}$$

and now we have only p and $P(e \mid \sim h)$ to consider. If we were to expand out $P(e \mid \sim h)$ we would find that it is a constant less than 1 multiplied by a sum whose terms are products $P(e \mid h_i)P(h_i)$,

where h_i are alternatives to h. Recall now the assumption, reasonably appropriate to the situation in 1915, that the only serious alternative to GTR was Classical Gravitation Theory (CGT), meaning that it is the only h_i apart from h itself such that $P(h_i)$ is not negligible. *Now* we bring in the additional assumption that P does not 'know' e. Judged on the residual background information alone, the fact that e is anomalous relative to CGT means therefore *that* $P(e \mid \sim h)$ *will be very small*, say ε.

We are now almost there, with just p itself to evaluate. Remember that this too is to be evaluated on the residual background information. Without any of the confirming evidence for h, including e, this should mean that p, though small by comparison with $P(\text{CGT})$, which is correspondingly large (e is now not an anomaly for CGT, since by assumption e does not exist), is not negligible. It follows that, because of the very large likelihood ratio in favour of h combined with a non-negligible if small prior probability, $P(h \mid e)$ is much larger than $p = P(h)$, and we see that h is correspondingly highly confirmed by e, *even though e is known*. The 'old evidence' problem is solved.

9.d | Conclusion

Our view, and we believe the only tenable view, of the Bayesian theory is of a theory of consistent probabilistic reasoning. Just as with the theory of deductive consistency, this gives rise automatically to an account of valid probabilistic inference, in which the truth, rationality, objectivity, cogency or whatever of the premises, here *prior probability assignments*, are exogenous considerations, just as they are in deductive logic. Not only are these features outside the scope of the theory, they are, for the reasons we have given, incapable of being given any coherent or sustainable interpretation in any case.

This is not to say that what we have presented here is the last word. Modesty alone would preclude this, but it is almost certainly anyway not true: the model of uncertain reasoning in this account is a crude and simple one, as crude and simple as the usual models of deductive inference. But it has also the explanatory strengths of these models which, crude as they are, still

dominate and mould discussions of deductive reasoning, and will continue to do so, in one version or another, for the foreseeable future. Which is saying a great deal. Enough, indeed, to end this book.

Bibliography

Akaike, H. 1973. Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium of Information Theory*, eds. B.N. Petrov and F. Csáki (Budapest: Akadémiai Kiadó), 267–281.

Anscombe, F.J. 1963. Sequential Medical Trials. *Journal of the American Statistical Association*, Volume 58, 365–383.

Anscombe, F.J., and R.J. Aumann. 1963. A Definition of Subjective Probability. *Annals of Mathematical Statistics*, Volume 34, 199–205.

Armitage, P. 1975. *Sequential Medical Trials.* Second edition. Oxford: Blackwell.

Atkinson, A.C. 1985. *Plots, Transformations, and Regression.* Oxford: Clarendon.

———. 1986. Comment: Aspects of Diagnostic Regression Analysis. *Statistical Science*, Volume 1, 397–402.

Babbage, C. 1827. Notice Respecting some Errors Common to many Tables of Logarithms. *Memoirs of the Astronomical* Society, Volume 3, 65–67.

Bacon, F. 1994 [1620]. *Novum Organum*. Translated and edited by P. Urbach and J. Gibson. Chicago: Open Court.

Barnett, V. 1973. *Comparative Statistical Inference.* New York: Wiley.

Bartha, P. 2004. Countable Additivity and the de Finetti Lottery. *British Journal for the Philosophy of Science*, Volume 55, 301–323.

Bayes, T. 1958 [1763]. An Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society*, Volume 53, 370–418. Reprinted with a biographical note by G.A. Barnard in *Biometrika* (1958), Volume 45, 293–315.

Belsley, D.A., E. Kuh, and R.E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York: Wiley.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii academiae scientiarum imperialis Petropolitanae*, Volume V, 175–192

Bernoulli, J. 1713. *Ars Conjectandi.* Basiliae.

Berry, D.A. 1989. Ethics and ECMO. *Statistical Science,* Volume 4, 306–310.

Blackwell, D., and L. Dubins. 1962. Merging of Opinions with Increasing Information. *Annals of Mathematical Statistics,* Volume 33, 882–87.

Bland, M. 1987. *An Introduction to Medical Statistics.* Oxford: Oxford University Press.

Blasco, A. 2001. The Bayesian Controversy in Animal Breeding. *Journal of Animal Science*, Volume 79, 2023–046.

Bovens, L. and S. Hartmann. 2003. *Bayesian Epistemology*. Oxford: Oxford University Press.

Bourke, G.J., L.E. Daly, and J. McGilvray. 1985. *Interpretation and Uses of Medical Statistics.* 3rd edition. St. Louis: Mosby.

Bowden, B.V. 1953. A Brief History of Computation. In *Faster than Thought,* edited by B.V. Bowden. London: Pitman.

Bradley, R. 1998. A Representation Theorem for a Decision Theory with Conditionals. *Synthese*, Volume 116, 187–229.

Brandt, R. 1986. 'Comment' on Chatterjee and Hadi (1986). *Statistical Science,* Volume 1, 405–07.

Broemeling, L.D. 1985. *Bayesian Analysis of Linear Models.* New York: Dekker.

Brook, R.J. and G.C. Arnold. 1985. *Applied Regression Analysis and Experimental Design.* New York: Dekker.

Burnham, K.P. and D.R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach.* New York: Springer-Verlag.

Byar, D.P. *et al.* (seven co-authors). 1976. Randomized Clinical Trials. *New England Journal of Medicine*, 74–80.

Byar, D.P. *et al.* (22 co-authors). 1990. Design Considerations for AIDS Trials. *New England Journal of Medicine*, Volume 323, 1343–48.

Carnap, R. 1947. On the Applications of Inductive Logic. *Philosophy and Phenomenological Research,* Volume 8, 133 148.

Casscells W., A. Schoenberger, and T. Grayboys. 1978. Interpretation by Physicians of Clinical Laboratory Results. *New England Journal of Medicine*, Volume 299, 999–1000.

Chatterjee, S., and A.S. Hadi. 1986. Influential Observations, High Leverage Points, and Outliers in Linear Regression. *Statistical Science,* Volume 1, 379–416.

Chatterjee, S., and B. Price. 1977. *Regression Analysis by Example.* New York: Wiley.

Chiang, C.L. 2003. *Statistical Methods of Analysis*. World Scientific Publishing.

Cochran, W.G. 1952. The χ^2 Test of Goodness of Fit. *Annals of Mathematical Statistics*, Volume 23, 315–345.

———. 1954. Some Methods for Strengthening the Common χ^2 Tests. *Biometrics,* Volume 10, 417–451.

Cook, R.D. 1986. Comment on Chatterjee and Hadi 1986. *Statistical Science,* Volume 1, 393–97.

Cournot, A.A. 1843. *Exposition de la Théorie des Chances et des Probabilités*. Paris.

Cox, D.R. 1968. Notes on Some Aspects of Regression Analysis. *Journal of the Royal Statistical Society,* Volume 131A, 265–279.

Cox, R.T. 1961. *The Algebra of Probable Inference.* Baltimore: The Johns Hopkins University Press.

Cramér, H. 1946. *Mathematical Methods of Statistics.* Princeton: Princeton University Press.

Daniel, C., and F.S. Wood. 1980. *Fitting Equations to Data.* New York: Wiley.

David, F.N. 1962. *Games, Gods, and Gambling.* London: Griffin.

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

Diaconis, P., and S.L. Zabell. 1982. Updating Subjective Probability. *Journal of the American Statistical Association,* Volume 77, 822–830.

Dobzhansky, T. 1967. Looking Back at Mendel's Discovery. *Science,* Volume 156, 1588–89.

Dorling, J. 1979. Bayesian Personalism, the Methodology of Research Programmes, and Duhem's Problem. *Studies in History and Philosophy of Science,* Volume 10, 177–187.

———. 1996. Further Illustrations of the Bayesian Solution of Duhem's Problem. http://www.princeton.edu/~bayesway/Dorling/dorling.html

Downham, J., ed. 1988. *Issues in Political Opinion Polling.* London: The Market Research Society. Occasional Papers on Market Research.

Duhem, P. 1905. *The Aim and Structure of Physical Theory.* Translated by P.P. Wiener, 1954. Princeton: Princeton University Press.

Dunn, J.M., and G. Hellman. 1986. Dualling: A Critique of an Argument of Popper and Miller. *British Journal for the Philosophy of Science,* Volume 37, 220–23.

Earman, J. 1992. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory.* Cambridge, Massachusetts: MIT Press.

Edwards, A.L. 1984. *An Introduction to Linear Regression and Correlation.* Second edition. New York: Freeman.

Edwards, A.W.F. 1972. *Likelihood.* Cambridge: Cambridge University Press.

———. 1986. Are Mendel's Results Really Too Close? *Biological Reviews of the Cambridge Philosophical Society,* Volume 61, 295–312.

Edwards, W. 1968. Conservatism in Human Information Processing. In *Formal Representation of Human Judgment,* B. Kleinmuntz, ed., 17–52.

Edwards, W., H. Lindman, and L.J. Savage. 1963. Bayesian Statistical Inference for Psychological Research. *Psychological Review,* Volume 70, 193–242.

Ehrenberg, A.S.C. 1975. *Data Reduction: Analysing and Interpreting Statistical Data.* London: Wiley.

FDA. 1988. *Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications.* Rockville: Center for Drug Evaluation and Research, Food and Drug Administration.

Feller, W. 1950. *An Introduction to Probability Theory and its Applications*, Volume 1. Third edition. New York: Wiley.

Feyerabend, P. 1975. *Against Method.* London: New Left Books.

Finetti, B. de. 1937. La prévision; ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré,* Volume 7, 1–68. Reprinted in 1964 in English translation as 'Foresight: Its Logical Laws, its Subjective Sources', in *Studies in Subjective Probability,* edited by H.E. Kyburg, Jr., and H.E. Smokler (New York: Wiley).

———. 1972. *Probability, Induction, and Statistics*, New York: Wiley.

———. 1974. *Theory of Probability*. Volume 1. New York: Wiley.

Fisher, R.A. 1922. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London,* Volume A222, 309–368.

———. 1930. Inverse Probability. *Proceedings of the Cambridge Philosophical Society*, Volume 26, 528–535.

———. 1935. Statistical Tests. *Nature,* Volume 136, 474.

———. 1936. Has Mendel's Work Been Rediscovered? *Annals of Science,* Volume 1, 115–137.

———. 1947 [1926]. *The Design of Experiments.* Fourth edition. Edinburgh: Oliver and Boyd.

———. 1956. *Statistical Methods and Statistical Inference.* Edinburgh: Oliver and Boyd.

———. 1970 [1925]. *Statistical Methods for Research Workers.* Fourteenth edition. Edinburgh: Oliver and Boyd.

Freeman, P.R. 1993. The Role of *P*-values in Analysing Trial Results. *Statistics in Medicine*, Volume 12, 1433–459.

Gabbay, D. 1994. What Is a Logical System? *What Is a Logical System?*, ed. D. Gabbay, Oxford: Oxford University Press, 179-217.

Gaifman, H. 1964. Concerning Measures in First Order Calculi. *Israel Journal of Mathematics*, Volume 2, 1–18.

———. 1979. Subjective Probability, Natural Predicates, and Hempel's Ravens. *Erkenntnis*, Volume 14, 105-159.

Gaifman, H., and M. Snir. Probabilities over Rich languages, Testing and Randomness. *Journal of Symbolic Logic* 47, 495–548.

Giere, R.N. 1984. *Understanding Scientific Reasoning.* Second edition. New York: Holt, Rinehart.

Gigerenzer, G. 1991. How to Make Cognitive Illusions Disappear: Beyond Heuristics and Biases. *European Review of Social Psychology,* Volume 3, 83–115.

Gillies, D.A. 1973. *An Objective Theory of Probability.* London: Methuen.

———. 1989. Non-Bayesian Confirmation Theory and the Principle of Explanatory Surplus. *Philosophy of Science Association 1988,* edited by A. Fine and J. Loplin, Volume 2 (Pittsburgh: Pittsburgh University Press), 373–381.

———. 1990. Bayesianism versus Falsificationism. *Ratio,* Volume 3, 82–98.

———. 2000. *Philosophical Theories of Probability*. London: Routledge.

Girotto, V., and M. Gonzalez. 2001. Solving Probabilistic and Statistical Problems: A Matter of Information Structure and Question Form. *Cognition*, Volume 78, 247-276.

Glymour, C. 1980. *Theory and Evidence.* Princeton: Princeton University Press.

Good, I.J. 1950. *Probability and the Weighing of Evidence.* London: Griffin.

———. 1961. The Paradox of Confirmation. *British Journal for the Philosophy of Science,* Volume 11, 63–64.

———. 1965. *The Estimation of Probabilities.* Cambridge, Massachusetts: MIT Press.

———. 1969. Discussion of Bruno de Finetti's Paper 'Initial Probabilities: A Prerequisite for any Valid Induction'. *Synthese,* Volume 20, 17–24.

———. 1981. Some Logic and History of Hypothesis Testing. In *Philosophical Foundations of Economics,* edited by J.C. Pitt (Dordrecht: Reidel).

———. 1983. Some History of the Hierarchical Bayes Methodology. *Good Thinking.* Minneapolis: University of Minnesota Press, 95–105.

Goodman, N. 1954. *Fact, Fiction, and Forecast.* London: Athlone.
Gore, S.M. 1981. Assessing Clinical Trials: Why Randomize? *British Medical Journal,* Volume 282, 1958–960.
Grünbaum, A. 1976. Is the Method of Bold Conjectures and Attempted Refutations *Justifiably* the Method of Science? *British Journal for the Philosophy of Science,* Volume 27, 105–136.
Gumbel, E.J. 1952. On the Reliability of the Classical Chi-Square Test. *Annals of Mathematical Statistics,* Volume 23, 253–263.
Gunst, R.F., and R.C. Mason. 1980. *Regression Analysis and its Application.* New York: Dekker.
Hacking, I. 1965. *Logic of Statistical Inference.* Cambridge: Cambridge University Press.
———. 1967. Slightly More Realistic Personal Probability. *Philosophy of Science,* Volume 34, 311–325.
———. 1975. *The Emergence of Probability.* Cambridge: Cambridge University Press.
Halmos, P. 1950. *Measure Theory.* New York: Van Nostrand.
Halpern, J.Y. 1999. Cox's Theorem Revisited. *Journal of Artificial Intelligence Research,* Volume 11, 429 435.
Hays, W.L. 1969 [1963]. *Statistics.* London: Holt, Rinehart and Winston.
Hays, W.L., and R.L. Winkler. 1970. *Statistics: Probability, Inference, and Decision,* Volume 1. New York: Holt, Rinehart.
Hellman, G. 1997. Bayes and Beyond. *Philosophy of Science,* Volume 64.
Hempel, C.G. 1945. Studies in the Logic of Confirmation. *Mind,* Volume 54, 1–26, 97–121. Reprinted in Hempel 1965.
———. 1965. *Aspects of Scientific Explanation.* New York: The Free Press.
———. 1966. *Philosophy of Natural Science.* Englewood Cliffs: Prentice-Hall.
Hodges, J.L., Jr., and E.L. Lehmann. 1970. *Basic Concepts of Probability and Statistics.* Second edition. San Francisco: Holden-Day.
Horwich, P. 1982. *Probability and Evidence.* Cambridge: Cambridge University Press.
———. 1984. Bayesianism and Support by Novel Facts. *British Journal for the Philosophy of Science,* Volume 35, 245–251.
Howson, C. 1973. Must the Logical Probability of Laws be Zero? *British Journal for the Philosophy of Science,* Volume 24, 153–163.
Howson, C., ed. 1976. *Method and Appraisal in the Physical Sciences.* Cambridge: Cambridge University Press.

———. 1987. Popper, Prior Probabilities, and Inductive Inference. *British Journal for the Philosophy of Science,* Volume 38, 207–224.
———. 1988a. On the Consistency of Jeffreys's Simplicity Postulate, and its Role in Bayesian Inference. *Philosophical Quarterly,* Volume 38, 68–83.
———. 1988b. Accommodation, Prediction, and Bayesian Confirmation Theory. *PSA 1988.* A. Fine and J. Leplin, eds., 381–392.
———. 1997. *Logic With Trees.* London: Routledge.
———. 2000. *Hume's Problem: Induction and the Justification of Belief.* Oxford: Clarendon.
———. 2002. Bayesianism in Statistics. *Bayes's Theorem,* ed. R. Swinburne, The Royal Academy: Oxford University Press, 39–71.
Hume, D. 1739. *A Treatise of Human Nature,* Books 1 and 2. London: Fontana.
———. 1777. *An Enquiry Concerning Human Understanding.* Edited by L.A. Selby-Bigge. Oxford: Clarendon.
Jaynes, E.T. 1968. Prior Probabilities. *Institute of Electrical and Electronic Engineers Transactions on Systems Science and Cybernetics,* SSC-4, 227–241.
———. 1973. The Well-Posed Problem. *Foundations of Physics,* Volume 3, 413–500.
———. 1983. *Papers on Probability, Statistics, and Statistical Physics,* edited by R. Rosenkrantz. Dordrecht: Reidel.
———. 1985. Some Random Observations. *Synthese,* Volume 63, 115–138.
———. 2003. *Probability Theory: The Logic of Science.* Cambridge: Cambridge University Press.
Jeffrey, R.C. 1970. 1983. *The Logic of Decision.* Second edition. Chicago: University of Chicago Press.
———. 2004. *Subjective Probability: The Real Thing.* Cambridge: Cambridge University Press.
Jeffreys, H. 1961. *Theory of Probability.* Third edition. Oxford: Clarendon.
Jennison, C., and B.W. Turnbull. 1990. Statistical Approaches to Interim Monitoring: A Review and Commentary. *Statistical Science,* Volume 5, 299–317.
Jevons, W.S. 1874. *The Principles of Science.* London: Macmillan.
Joyce, J.M. 1998. A Nonpragmatic Vindication of Probabilism. *Philosophy of Science,* Volume 65, 575–603.
———. 1999. *The Foundations of Causal Decision Theory.* Cambridge: Cambridge University Press.

Kadane, J., *et al.* 1980. Interactive Elicitation of Opinion for a Normal Linear Model. *Journal of the American Statistical Association,* Volume 75, 845–854.

Kadane, J.B., M.J. Schervish, and T. Seidenfeld. 1999. *Rethinking the Foundations of Statistics*. Cambridge: Cambridge University Press.

Kadane, J.B. and T. Seidenfeld. 1990. Randomization in a Bayesian Perspective. *Journal of Statistical Planning and Inference*, Volume 25, 329–345.

Kant, I. 1783. *Prolegomena to any Future Metaphysics*. Edited by L.W. Beck, 1950. Indianapolis: Bobbs-Merrill.

Kempthorne, O. 1966. Some Aspects of Experimental Inference. *Journal of the American Statistical Association,* Volume 61, 11–34.

———. 1971. Probability, Statistics, and the Knowledge Business. In *Foundations of Statistical Inference,* edited by V.P. Godambe and D.A. Sprott. Toronto: Holt, Rinehart and Winston of Canada.

———. 1979. *The Design and Analysis of Experiments.* Huntington: Robert E. Krieger.

Kendall, M.G., and A. Stuart. 1979. *The Advanced Theory of Statistics,* Volume 2. Fourth edition. London: Griffin.

———. 1983. *The Advanced Theory of Statistics,* Volume 3. Fourth edition. London: Griffin.

Keynes, J.M. 1921. *A Treatise on Probability.* London: Macmillan.

Kieseppä, I.A. 1997. Akaike Information Criterion, Curve-fitting, and the Philosophical Problem of Simplicity. *British Journal for the Philosophy of Science*, Volume 48, 21–48.

Kitcher, P. 1985. *Vaulting Ambition.* Cambridge, Massachusetts: MIT Press.

Kolmogorov, A.N. 1950. *Foundations of the Theory of Probability.* Translated from the German of 1933 by N. Morrison. New York: Chelsea Publishing. Page references are to the 1950 edition.

Korb, K.B. 1994. Infinitely Many Resolutions of Hempel's Paradox. In *Theoretical Aspects of Reasoning about Knowledge*, 138–49, edited by R. Fagin. Asilomar: Morgan Kaufmann.

Kuhn, T.S. 1970 [1962]. *The Structure of Scientific Revolutions.* Second edition. Chicago: University of Chicago Press.

Kyburg, H.E., Jr., and E. Smokler, eds. 1980. *Studies in Subjective Probability.* Huntington: Krieger.

Lakatos, I. 1963. Proofs and Refutations. *British Journal for the Philosophy of Science*, Volume 14, 1–25, 120–139, 221–143, 296, 432.

———. 1968. Criticism and the Methodology of Scientific Research Programmes. *Proceedings of the Aristotelian Society,* Volume 69, 149–186.

———. 1970. Falsification and the Methodology of Scientific Research Programmes. In *Criticism and the Growth of Knowledge,* edited by I. Lakatos and A. Musgrave. Cambridge: Cambridge University Press.

———. 1974. Popper on Demarcation and Induction. In *The Philosophy of Karl Popper*, edited by P.A. Schilpp. La Salle: Open Court.

———. 1978. *Philosophical Papers.* Two volumes. Edited by J. Worrall and G. Currie. Cambridge: Cambridge University Press.

Laplace, P.S. de. 1820. *Essai Philosophique sur les Probabilités.* Page references are to *Philosophical Essay on Probabilities,* 1951. New York: Dover.

Lee, P.M. 1997. *Bayesian Statistics*. Second edition. London: Arnold.

Lewis, D. 1981. A Subjectivist's Guide to Objective Chance. In *Studies in Inductive Logic and Probability,* edited by R.C. Jeffrey, 263–293. Berkeley: University of California Press.

Lewis-Beck, M.S. 1980. *Applied Regression.* Beverley Hills: Sage.

Li, M. and P.B.M. Vitanyi. 1997. *An Introduction to Kolmogorov Complexity Theory and its Applications*. Second edition. Berlin: Springer.

Lindgren, B.W. 1976. *Statistical Theory.* Third edition. New York: Macmillan.

Lindley, D.V. 1957. A Statistical Paradox. *Biometrika,* Volume 44, 187–192.

———. 1965. *Introduction to Probability and Statistics, from a Bayesian Viewpoint.* Two volumes. Cambridge: Cambridge University Press.

———. 1970. Bayesian Analysis in Regression Problems. In *Bayesian Statistics,* edited by D.L. Meyer and R.O. Collier. Itasca: F.E. Peacock.

———. 1971. *Bayesian Statistics: A Review*. Philadelphia: Society for Industrial and Applied Mathematics.

———. 1982. The Role of Randomization in Inference. *Philosophy of Science Association,* Volume 2, 431–446.

———. 1985. *Making Decisions.* Second edition. London: Wiley.

Lindley, D.V., and G.M. El-Sayyad. 1968. The Bayesian Estimation of a Linear Functional Relationship. *Journal of the Royal Statistical Society,* Volume 30B, 190–202.

Lindley, D.V., and L.D. Phillips. 1976. Inference for a Bernoulli Process (a Bayesian View). *American Statistician,* Volume 30, 112–19.

Mackie, J.L. 1963. The Paradox of Confirmation. *British Journal for the Philosophy of Science,* Volume 38, 265–277.

McIntyre, I.M.C. 1991. Tribulations for Clinical Trials. *British Medical Journal,* Volume 302, 1099–1100.

Maher, P. 1990. Why Scientists Gather Evidence. *British Journal for the Philosophy of Science,* Volume 41, 103–119.

——— 1990. Acceptance Without Belief. *PSA 1990*, Volume 1, eds. A. Fine, M. Forbes, and L. Wessels, 381–392.

———1997. Depragmatized Dutch Book Arguments. *Philosophy of Science*, Volume 64, 291–305.

Mallet, J.W. 1880. Revision of the Atomic Weight of Aluminium. *Philosophical Transactions,* Volume 171, 1003–035.

———. 1893. The Stas Memorial Lecture. In *Memorial Lectures delivered before the Chemical Society 1893-1900.* Published 1901. London. Gurney and Jackson.

Mann, H.B., and A. Wald. 1942. On the Choice of the Number of Intervals in the Application of the Chi-Square Test. *Annals of Mathematical Statistics,* Volume 13, 306–317.

Mayo, D.G. 1996. *Error and the Growth of Experimental Knowledge.* Chicago: University of Chicago Press.

Medawar, P. 1974. More Unequal than Others. *New Statesman,* Volume 87, 50–51.

Meier, P. 1975. Statistics and Medical Experimentation. *Biometrics,* Volume 31, 511–529.

Miller, D. 1991. On the Maximization of Expected Futility. PPE Lectures, Lecture 8. Department of Economics: University of Vienna.

Miller, R. 1987. *Bare-faced Messiah.* London: Michael Joseph.

Mises, R. von. 1939 [1928]. *Probability, Statistics, and Truth.* First English edition prepared by H. Geiringer. London: Allen and Unwin.

———. 1957. Second English edition, revised, of *Probability, Statistics and Truth.*

———. 1964. *Mathematical Theory of Probability and Statistics.* New York: Academic Press.

Mood, A.M. 1950. *Introduction to the Theory of Statistics.* New York: McGraw-Hill.

Mood, A.M., and F.A. Graybill. 1963. *Introduction to the Theory of Statistics.* New York: McGraw-Hill.

Musgrave, A. 1975. Popper and 'Diminishing Returns from Repeated Tests', *Australasian Journal of Philosophy,* Volume 53, 248–253.

Myrvold, W.C. and W.L. Harper. 2002. Model Selection and Scientific Inference. *Philosophy of Science*, Volume 69, S124–134.

Neyman, J. 1935. On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection. Reprinted in Neyman 1967, 98–141.

———. 1937. Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society,* Volume 236A, 333–380.

———. 1941. Fiducial Argument and the Theory of Confidence Intervals. *Biometrika,* Volume 32, 128–150. Page references are to the reprint in Neyman 1967.

———. 1952. *Lectures and Conferences on Mathematical Statistics and Probability.* Second edition. Washington, D.C.: U.S. Department of Agriculture.

———. 1967. *A Selection of Early Statistical Papers of J. Neyman.* Cambridge: Cambridge University Press.

Neyman, J., and E.S. Pearson. 1928. On the Use and the Interpretation of Certain Test Criteria for Purposes of Statistical Inference. *Biometrika,* Volume 20, 175–240 (Part I), 263–294 (Part II).

———. 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society,* Volume 231A, 289–337. Page references are to the reprint in Neyman and Pearson's *Joint Statistical Papers* (Cambridge: Cambridge University Press, 1967).

Pais, A. 1982. *Subtle Is the Lord.* Oxford: Clarendon.

Paris, J. 1994. *The Uncertain Reasoner's Companion.* Cambridge: Cambridge University Press.

Paris, J. and A. Vencovskà. 2001. Common Sense and Stochastic Independence. *Foundations of Bayesianism*, eds. D. Corfield and J. Williamson. Dordrecht: Kluwer, 203–241.

Pearson, E.S. 1966. Some Thoughts on Statistical Inference. In *The Selected Papers of E.S. Pearson,* 276–183. Cambridge: Cambridge University Press.

Pearson, K. 1892. *The Grammar of Science.* Page references are to the edition of 1937 (London: Dent).

Peto, R., *et al.* 1988. Randomised Trial of Prophylactic Daily Aspirin in British Male Doctors. *British Medical Journal,* Volume 296, 313–331.

Phillips, L.D. 1973. *Bayesian Statistics for Social Scientists.* London: Nelson.

———. 1983. A Theoretical Perspective on Heuristics and Biases in Probabilistic Thinking. In *Analysing and Aiding Decision,* edited by P.C. Humphreys, O. Svenson, and A. Van. Amsterdam: North Holland.

Pitowsky, I. 1994. George Boole's Conditions of Possible Experience and the Quantum Puzzle. *British Journal for the Philosophy of Science*, Volume 45, 95–127.

Poincaré, H. 1905. *Science and Hypothesis*. Page references are to the edition of 1952 (New York: Dover).

Polanyi, M. 1962. *Personal Knowledge*. Second edition. London: Routledge.

Pollard, W. 1985. *Bayesian Statistics for Evaluation Research: An Introduction*. Beverly Hills: Sage.

Polya, G. 1954. *Mathematics and Plausible Reasoning*, Volumes 1 and 2. Princeton: Princeton University Press.

Popper, K.R. 1959. The Propensity Interpretation of Probability. *British Journal for the Philosophy of Science*, Volume 10, 25–42.

———. 1959a. *The Logic of Scientific Discovery*. London: Hutchinson.

———. 1960. *The Poverty of Historicism*. London: Routledge.

———. 1963. *Conjectures and Refutations*. London: Routledge.

———. 1972. *Objective Knowledge*. Oxford: Oxford University Press.

———. 1983. A Proof of the Impossibility of Inductive Probability. *Nature*, Volume 302, 687–88.

Pratt, J.W. 1962. On the Foundations of Statistical Inference. *Journal of the American Statistical Association*, Volume 57, 269–326.

———. 1965. Bayesian Interpretation of Standard Inference Statements. *Journal of the Royal Statistical Society*, 27B, 169–203.

Pratt, J.W., H. Raiffa, and R. Schlaifer. 1965. *Introduction to Statistical Decision Theory*.

Prout, W. 1815. On the Relation Between the Specific Gravities of Bodies in Their Gaseous State and the Weights of Their Atoms. *Annals of Philosophy*, Volume 6, 321–330. Reprinted in *Alembic Club Reprints*, No. 20, 1932, 25–37 (Edinburgh: Oliver and Boyd).

Prout, W. 1816. Correction of a Mistake in the Essay on the Relations Between the Specific Gravities of Bodies in Their Gaseous State and the Weights of their Atoms. *Annals of Philosophy*, Volume 7, 111–13.

Putnam, H. 1975. *Collected Papers*, Volume 2. Cambridge: Cambridge University Press.

Ramsey, F.P. 1931. Truth and Probability. In Ramsey, *The Foundations of Mathematics and Other Logical Essays* (London: Routledge).

Rao, C.D. 1965. *Linear Statistical Inference and its Applications*. New York: Wiley.

Renyí, A. 1955. On a New Axiomatic Theory of Probability. *Acta Mathematica Academiae Scientiarum Hungaricae*, Volume VI, 285–335.

Rosenkrantz, R.D. 1977. *Inference, Method, and Decision: Towards a Bayesian Philosophy of Science*. Dordrecht: Reidel.

Salmon, W.C. 1981. Rational Prediction. *British Journal for the Philosophy of Science*, Volume 32, 115–125.

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

———. 1962. Subjective Probability and Statistical Practice. In *The Foundations of Statistical Inference*, edited by G.A. Barnard and D.R. Cox (New York: Wiley), 9–35.

———. 1962a. A Prepared Contribution to the Discussion of Savage 1962, 88–89, in the same volume.

Schervish, M., T. Seidenfeld, and J.B. Kadane. 1990. State-Dependent Utilities. *Journal of the American Statistical Association*, Volume 85, 840–847.

Schroeder, L.D., D.L. Sjoquist, and P.E. Stephan. 1986. *Understanding Regression Analysis*. Beverly Hills: Sage.

Schwarz, G. 1978. Estimating the Dimension of a Model. *Annals of Statistics*, Volume 6, 461–464.

Schwartz, D., R. Flamant and J. Lellouch. 1980. *Clinical Trials [L'essay therapeutique chez l'homme]*. New York: Academic Press. Translated by M.J.R. Healy.

Scott, D. and P. Krauss. 1966. Assigning Probabilities to Logical Formulas. *Aspects of Inductive Logic*, eds. J. Hintikka and P. Suppes. Amsterdam: North Holland, 219–264.

Seal, H.L. 1967. The Historical Development of the Gauss Linear Model. *Biometrika*, Volume 57, 1–24.

Seber, G.A.F. 1977. *Linear Regression Analysis*. New York: Wiley.

Seidenfeld, T. 1979. *Philosophical Problems of Statistical Inference*. Dordrecht: Reidel.

———. 1979. Why I Am Not an Objective Bayesian: Some Reflections Prompted by Rosenkrantz. *Theory and Decision*, Volume 11, 413–440.

Shimony, A. 1970. Scientific Inference. In *Pittsburgh Studies in the Philosophy of Science*, Volume 4, edited by R.G. Colodny. Pittsburgh: Pittsburgh University Press.

———. 1985. The Status of the Principle of Maximum Entropy. *Synthese*, Volume 68, 35–53.

———. 1993 [1988]. An Adamite Derivation of the Principles of the Calculus of Probability. In Shimony, *The Search for a Naturalistic World View*, Volume 1 (Cambridge: Cambridge University Press), 151–162.

Shore, J.E. and R.W. Johnson. 1980. Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. *IEEE Transactions on Information Theory* 26:1, 26–37.

Skyrms, B. 1977. *Choice and Chance*. Belmont: Wadsworth.

Smart, W.M. 1947. John Couch Adams and the Discovery of Neptune. *Occasional Notes of the Royal Astronomical Society*, No. 11.

Smith, T.M. 1983. On the Validity of Inferences from Non-random Samples. *Journal of the Royal Statistical Society,* Volume 146A, 394–403.

Smullyan, R. 1968. *First Order Logic*. Berlin: Springer.

Sober, E. and M. Forster. 1994. How to Tell When Simpler, More Unified, Or Less Ad Hoc Theories Will Provide More Accurate Predictions. *British Journal for the Philosophy of Science*, Volume 45, 1–37.

Spielman, S. 1976. Exchangeability and the Certainty of Objective Randomness. *Journal of Philosophical Logic,* Volume 5, 399–406.

Sprent, P.1969. *Models in Regression.* London: Methuen.

Sprott, W.J.H. 1936. Review of K. Lewin's *A Dynamical Theory of Personality*. *Mind,* Volume 45, 246–251.

Stachel, J. 1998. *Einstein's Miraculous Year: Five Papers that Changed the Face of Physics*. Princeton: Princeton University Press.

Stas, J.S. 1860. Researches on the Mutual Relations of Atomic Weights. *Bulletin de l'Académie Royale de Belgique,* 208–336. Reprinted in part in *Alembic Club Reprints,* No. 20, 1932 (Edinburgh: Oliver and Boyd), 41–47.

Stuart, A 1954. Too Good to Be True. *Applied Statistics,* Volume 3, 29–32.

———. 1962. *Basic Ideas of Scientific Sampling.* London: Griffin.

Sudbery, A. 1986. *Quantum Mechanics and the Particles of Nature.* Cambridge: Cambridge University Press.

Suzuki, S. 2005. The Old Evidence Problem and AGM Theory. *Annals of the Japan Association for Philosophy of Science*, 1–20.

Swinburne, R.G. 1971. The Paradoxes of Confirmation: A Survey. *American Philosophical Quarterly,* Volume 8, 318–329.

Tanur, J.M., *et al.* 1989. *Statistics: A Guide to the Unknown*. Third Edition. Duxbury Press.

Teller, P. 1973. Conditionalisation and Observation. *Synthese,* Volume 26, 218–258.

Thomson, T. 1818. Some Additional Observations on the Weights of the Atoms of Chemical Bodies. *Annals of Philosophy,* Volume 12, 338–350.

Uffink, J. 1995. Can the Maximum Entropy Method be Explained as a Consistency Requirement? *Studies in the History and Philosophy of Modern Physics*, Volume 26B, 223–261.

Urbach, P. 1981. On the Utility of Repeating the 'Same Experiment'. *Australasian Journal of Philosophy,* Volume 59, 151–162.

———. 1985. Randomization and the Design of Experiments. *Philosophy of Science,* Volume 52, 256–273.

———. 1987. *Francis Bacon's Philosophy of Science.* La Salle: Open Court.

———. 1987a. Clinical Trial and Random Error. *New Scientist,* Volume 116, 52–55.

———. 1987b. The Scientific Standing of Evolutionary Theories of Society. *The LSE Quarterly,* Volume 1, 23–42.

———. 1989. Random Sampling and the Principles of Estimation. *Proceedings of the Aristotelian Society,* Volume 89, 143–164.

———. 1991. Bayesian Methodology: Some Criticisms Answered. *Ratio (New Series),* Volume 4, 170–184.

———. 1992. Regression Analysis: Classical and Bayesian. *British Journal for the Philosophy of Science,* Volume 43, 311–342.

———. 1993. The Value of Randomization and Control in Clinical Trials. *Statistics in Medicine,* Volume 12, 1421–431.

Van Fraassen, B.C. 1980. *The Scientific Image.* Oxford: Clarendon.

———. 1983. Calibration: A Frequency Justification for Personal Probability. In R.S. Cohen and L. Laudan, eds., *Physics, Philosophy, and Psychoanalysis* (Dordrecht: Reidel), 295–321.

———. 1984. Belief and the Will. *Journal of Philosophy,* Volume LXXXI, 235–256.

———. 1989. *Laws and Symmetry.* Oxford: Clarendon.

Velikovsky, I. 1950. *Worlds in Collision.* London: Gollancz. Page references are to the 1972 edition, published by Sphere.

Velleman, P.F. 1986. Comment on Chatterjee, S., and Hadi, A.S. 1986. *Statistical Science,* Volume 1, 412–15.

Velleman, P.F., and R.E. Welsch. 1981. Efficient Computing of Regression Diagnostics. *American Statistician,* Volume 35, 234–242.

Venn, J. 1866. *The Logic of Chance.* London: Macmillan.

Vranas, P.B.M. 2004. Hempel's Raven Paradox: A Lacuna in the Standard Bayesian Solution. *British Journal for the Philosophy of Science*, Volume 55, 545–560.

Wall, P. 1999. *Pain: The Science of Suffering*. London: Weidenfeld and Nicolson.

Watkins, J.W.N. 1985. *Science and Scepticism.* London: Hutchinson and Princeton: Princeton University Press.

———. 1987. A New View of Scientific Rationality. In *Rational Change in Science,* edited by J. Pitt and M. Pera. Dordrecht: Reidel.

Weinberg, S. and K. Goldberg. 1990. *Statistics for the Behavioral Sciences*. Cambridge: Cambridge University Press.

Weisberg, S. 1980. *Applied Linear Regression.* New York: Wiley.

Welsch, R.E. 1986. Comment on Chatterjee, S., and Hadi, A.S. 1986. *Statistical Science,* Volume 1, 403–05.

Whitehead, J. 1993. The Case for Frequentism in Clinical Trials. *Statistics in Medicine,* Volume 12, 1405–413.

Williams, P.M. 1980. Bayesian Conditionalisation and the Principle of Minimum Information. *British Journal for the Philosophy of Science,* Volume 31, 131–144.

Williamson, J. 1999. Countable Additivity and Subjective Probability. *British Journal for the Philosophy of Science*, Volume 50, 401–416.

Williamson, J. 2005. *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford: Oxford University Press.

Williamson, J. and D. Corfield. 2001. Introduction: Bayesianism into the Twenty-First Century. In Corfield, D. and Williamson, J., eds., *Foundations of Bayesianism* (Dordrecht: Kluwer).

Wonnacott, T.H., and R.J. Wonnacott. 1980. *Regression: A Second Course in Statistics*. New York: Wiley.

Wood, M. 2003. *Making Sense of Statistics*. New York: Palgrave Macmillan.

Yates, F. 1981. *Sampling Methods for Censuses and Surveys.* Fourth edition. London: Griffin.

Index