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.

Copyright © 2006 by Carus Publishing Company

First printing 2006

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, Open Court Publishing Company, a division of Carus Publishing Company, 315 Fifth Street, P.O. Box 300, Peru, Illinois 61354-0300.

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

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 **1** of the algebra, and all logical falsehoods map to the unique least element **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)** $P(a \mathbf{v} a) = P(a) + P(\sim a)$. But by **(2)** $P(a \mathbf{v} \sim a) = 1$, 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 $a \mathbf{v} \sim b$ is a logical truth if $a \Leftrightarrow b$.
Assume that a $\Leftrightarrow b$. Then $P(a \text{ v } \sim b) = 1$. Also if a $\Leftrightarrow b$ then a entails $\sim\sim b$ so $P(a \mathbf{v} \sim b) = P(a) + P(\sim b)$.
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 $[a \mathbf{v} (b \,\&\, \sim a)] \Leftrightarrow b$. Hence by **(7)** $P(b) = P[a \mathbf{v} (b \,\&\, \sim a)]$. But a entails $\sim(b \,\&\, \sim a)$ and so $P[a \mathbf{v} (b \,\&\, \sim a)] = P(a) + P(b \,\&\, \sim a)$. 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 $P(a_1 \mathbf{v} \ldots \mathbf{v} a_n) = 1$ 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 $P(a_1 \mathbf{v} \ldots \mathbf{v} a_n) = 1$ 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) $Q(a \mathbf{v} c) = Q(a) + Q(c)$, 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:

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

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* $P(h_1 \mathbf{v} \ldots \mathbf{v} h_n) = 1$ 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 $h_1 \mathbf{v} \ldots \mathbf{v} h_n$ 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 math-

ematical 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 infi-

nite 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 $\Sigma P(a \,\&\, b_j)$. 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.

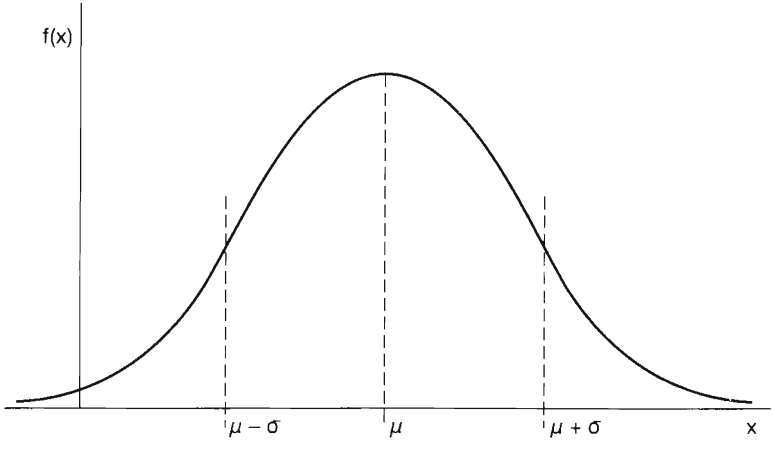

FIGURE 2.1

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) $P(X < x \mid y < Y < y + \delta y) = \dfrac{P(X < x \;\&\; y < Y \leq y + \delta y)}{P(y < Y \leq y + \delta y)}$.

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 $\dfrac{f(x,y)}{f(y)}$ 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

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

2.1 | 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

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

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.

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.