Scientific Method in Brief

Hugh G. Gauch, Jr.
Cornell University, New York

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9781107666726

First published 2012

Printed and Bound in the United Kingdom by the MPG Books Group

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
Gauch, Jr., Hugh G., 1942–
Scientific method in brief / Hugh G. Gauch, Jr.
p. cm.
ISBN 978-1-107-66672-6 (pbk.)
1. Science – Methodology. I. Title.
Q175.G3368 2012
001.4′2 – dc23 2012016520

ISBN 978-1-107-66672-6 Paperback

9

Inductive logic and statistics

The logic that is so essential for scientific reasoning, being the "L" portion of the PEL model, is of two basic kinds: deductive and inductive. Chapter 7 reviewed deductive logic, and Chapter 8 probability, which is a branch of deductive logic. This chapter reviews inductive logic, with "statistics" being essentially the term meaning applied inductive logic.

A considerable complication is that statisticians have two competing paradigms for induction: Bayesian and frequentist statistics. At stake are scientific concerns, seeking efficient extraction of information from data to answer important questions, and philosophical concerns, involving rational foundations and coherent reasoning.

This chapter cannot possibly do what entire books on statistics do – present a comprehensive treatment. But it can provide a prolegomenon to clarify the most basic and pivotal issues, which are precisely the aspects of statistics that scientists generally comprehend the least. The main objectives are to depict and contrast the Bayesian and frequentist paradigms and to explain why inductive logic or statistics often functions well despite imperfect data, imperfect models, and imperfect scientists. Extremely important research in agriculture, medicine, engineering, and other fields imposes great responsibilities on statistical practice.

Historical perspective on induction

This section gives a brief history of induction from Aristotle to John Stuart Mill, with more recent developments deferred to later sections. Aristotle (384–322 BC) had a broad conception of induction. Primarily, induction is reasoning from particular instances to general conclusions. That is ampliative reasoning from observed to unobserved, from part to whole, from sample to population. Aristotle cautioned against hasty generalizations and noted that a single counter example suffices to nullify a universal generalization. He carefully distinguished induction from deduction, analogy, and isolated examples.

One of Aristotle's most influential contributions to the philosophy of science was his model of scientific logic or reasoning, the inductive–deductive method. Scientific inquiry alternates inductive and deductive steps. From observations, induction provides general principles, and with those principles serving as premises, deduction predicts or explains observed phenomena. Overall, there is an advance from knowledge of facts to knowledge of an explanation for the facts.

Epicurus (341–271 BC) discussed the fundamental role of induction in forming concepts and learning language in his doctrine of "anticipation." From repeated sense perceptions, a general idea or image is formed that combines the salient, common features of the objects, such as the concept of a horse derived from numerous observations of horses. Once stored in memory, this concept or anticipation acts as an organizing principle or convention for discriminating which perceptions or objects are horses and for stating truths about horses.

Robert Grosseteste (*c.* 1168–1253) affirmed and refined Aristotle's inductive–deductive method, which he termed the Method of Resolution and Composition for its inductive and deductive components, respectively. But he added to Aristotle's methods of induction. His purposes were to verify true theories and to falsify false theories. Causal laws were suspected when certain phenomena were frequently correlated, but natural science sought robust knowledge of real causes, not accidental correlations. "Grosseteste's contribution was to emphasize the importance of *falsification* in the search for true causes and to develop the method of verification and falsification into a systematic method of experimental procedure" (Crombie 1962:84). His approach used deduction to falsify proposed but defective inductions. As mentioned in the earlier chapter on deduction, Grosseteste's Method of Verification deduced consequences of a theory beyond its original application and then checked those predictions experimentally. His Method of Falsification eliminated bad theories by deducing implications known to be false.

Grosseteste clearly understood that his optimistic view of induction required two metaphysical presuppositions about the nature of physical reality: the uniformity of nature and the principle of parsimony or simplicity. Without those presuppositions, there is no defensible method of induction in particular or method of science in general.

In essence, at Oxford in 1230, Grosseteste's new scientific method – with its experiments, Method of Resolution and Composition, Method of Verification, Method of Falsification, emphasis on logic and parsimony, and commonsense presuppositions – was the paradigm for the design and analysis of scientific experiments. Science's goal was to provide humans with truth about the physical world, and induction was a critical component of scientific method.

Roger Bacon (*c.* 1214–1294) promulgated three prerogatives of experimental science, as mentioned in Chapter 3. Of those, the first two concerned induction. His first prerogative was that inductive conclusions should be submitted to

further testing. That was much like his predecessor Grosseteste's Method of Verification. His second prerogative was that experiments could increase the amount and variety of data used by inductive inferences, thus helping scientists to discriminate between competing hypotheses.

John Duns Scotus (*c.* 1265–1308), at Paris, reflected Oxford's confidence about inductive logic. He admired Grosseteste's commentaries on Aristotle's *Posterior Analytics* and *Physics* but disagreed on some points. Duns Scotus admitted that, ordinarily, induction could not reach evident and certain knowledge through complete enumeration, and yet he was quite optimistic that "probable knowledge could be reached by induction from a sample and, moreover, that the number of instances observed of particular events being correlated increased the probability of the connexion between them being a truly universal and causal one. . . . He realized that it was often impossible to get beyond mere empirical generalizations, but he held that a well-established empirical generalization could be held with certainty because of the principle of the uniformity of nature, which he regarded as a self-evident assumption of inductive science" (Crombie 1962:168–169).

Building on an earlier proposal by Grosseteste, Duns Scotus offered an inductive procedure called the Method of Agreement. "The procedure is to list the various circumstances that are present each time the effect occurs, and to look for some one circumstance that is present in every instance" (Losee 2001:29–30). For example, if circumstances *ABCD*, *ACE*, *ABEF*, and *ADF* all gave rise to the same effect *x*, then one could conclude that *A* could be the cause of *x*, although Duns Scotus cautiously refrained from the stronger claim that *A* must be the cause of *x*. The Method of Agreement could promote scientific advances by generating plausible hypotheses that merited further research to reach a more nearly definitive conclusion.

Henry of Ghent (*c.* 1217–1293), in contrast to Duns Scotus, believed that real knowledge had to be about logically necessary things, not the contingent things of which the physical world is composed. Had his view prevailed, science in general and induction in particular would now be held in low philosophical esteem.

William of Ockham (*c.* 1285–1347) further developed inductive logic along lines begun earlier by Grosseteste and Duns Scotus. He added another inductive procedure, the Method of Difference. "Ockham's method is to compare two instances – one instance in which the effect is present, and a second instance in which the effect is not present" (Losee 2001:30–31). For example, if circumstances *ABC* gave effect *x*, but circumstances *AB* did not, then one could conclude that *C* could be the cause of *x*. But Ockham was cautious in such claims, especially because he realized the difficulty in proving that two cases differed in only one respect. As a helpful, although partial, solution, he recommended comparing a large number of cases to reduce the possibility that an unrecognized factor could be responsible for the observed effect *x*.

Nicholas of Autrecourt (*c.* 1300–1350) had the most skeptical view of induction among medieval thinkers, prefiguring the severe challenge that would come several centuries later from David Hume. "He insisted that it cannot be established that a correlation which has been observed to hold must continue to hold in the future" (Losee 2001:37). Indeed, if the uniformity of nature is questioned in earnest, then induction is in big trouble. Recall that Grosseteste had recognized that induction depended on the uniformity of nature.

Sir Francis Bacon (1561–1626) so emphasized induction that his conception of scientific method is often known as Baconian induction. He criticized Aristotelian induction on three counts: haphazard data collection without systematic experimentation; hasty generalizations, often later proved false; and simplistic enumerations, with inadequate attention to negative instances.

Bacon discussed two inductive methods. The old and defective procedure was the "anticipation of nature," with "anticipation" reflecting its Epicurean usage, which led to hasty and frivolous inductions. The new and correct procedure was the "interpretation of nature." Inductions or theories that were acceptable interpretations "must encompass more particulars than those which they were originally designed to explain and, secondly, some of these new particulars should be verified," that is, "theories must be larger and wider than the facts from which they are drawn" (Urbach 1987:28). Good inductive theories would have predictive success.

René Descartes (1596–1650) deemed Bacon's view untenable, so he attempted to invert Bacon's scientific method: "But whereas Bacon sought to discover general laws by progressive inductive ascent from less general relations, Descartes sought to begin at the apex and work as far downwards as possible by a deductive procedure" (Losee 2001:64). Of course, that inverted strategy shifted the burden to establishing science's first principles, which had its own challenges.

Sir Isaac Newton (1642–1727) developed an influential view of scientific method that was directed against Descartes's attempt to derive physical laws from metaphysical principles. Rather, Newton insisted on careful observation and induction, saying that "although the arguing from Experiments and Observations by Induction be no Demonstration of general Conclusions, yet it is the best way of arguing which the Nature of Things admits of" (Losee 2001:73). Newton affirmed Aristotle's inductive–deductive method, which Newton termed the "Method of Analysis and Synthesis" for its deductive and inductive components, respectively. "By insisting that scientific procedure should include both an inductive stage and a deductive stage, Newton affirmed a position that had been defended by Grosseteste and Roger Bacon in the thirteenth century, as well as by Galileo and Francis Bacon at the beginning of the seventeenth century" (Losee 2001:73).

In Newton's scientific method, induction was extremely prominent, being no less than one of his four rules of scientific reasoning: "In experimental philosophy we are to look upon propositions collected by general induction from

phænomena as accurately or very nearly true, notwithstanding any contrary hypotheses that may be imagined, till such time as other phænomena occur, by which they may either be made more accurate, or liable to exceptions" (Williams and Steffens 1978:286).

John Stuart Mill (1806–1873) wrote a monumental *System of Logic* that covered deductive and inductive logic, with a subtitle proclaiming a connected view of the principles of evidence and the methods of scientific investigation. Like Francis Bacon, Mill recommended a stepwise inductive ascent from detailed observations to general theories. He had four (or five) inductive methods for discovering scientific theories or laws that were essentially the same as those of Grosseteste, Duns Scotus, and Ockham. Despite his enthusiasm for induction, Mill recognized that his methods could not work well in cases of multiple causes working together to produce a given effect. Mill wanted not merely to discover scientific laws but also to justify and prove them, while carefully distinguishing real causal connections from merely accidental sequences. But his justification of induction has not satisfied subsequent philosophers of science.

More recent developments in inductive logic will be discussed later in this chapter. During the twentieth century, induction picked up a common synonym: statistics. Statistics *is* inductive logic. The historically recent advent of statistical methods, digital computers, and enormous databases has stimulated and facilitated astonishing advances in induction.

Bayesian inference

For a simple example of Bayesian inference about which hypothesis is true, envision joining an introductory statistics class as they perform an experiment. The professor shows the class an ordinary fair coin, an opaque urn, and some marbles identical except for color, being either blue or white. Two volunteers, students Juan and Beth, are appointed as experimentalists. Juan receives his instructions and executes the following: He flips the coin without showing it to anyone else. If the coin toss gives heads, he is to place in the urn one white marble and three blue marbles. But if the coin toss gives tails, he is to place in the urn three white marbles and one blue marble. Juan knows the urn's contents, but the remainder of the class, including Beth and the professor, know only that exactly one of two hypotheses is true: either H_B, that the urn contains one white marble and three blue marbles, or else H_W, that it contains three white marbles and one blue marble.

The class is to determine which hypothesis, H_B or H_W, is probably true, by means of the following experiment: Beth is to mix the marbles, draw one marble, show its color to the class, and then replace it in the urn. That procedure is to be repeated as necessary. The stopping rule is to stop when either hypothesis reaches or exceeds a probability of 0.999. In other words, there is to be at most

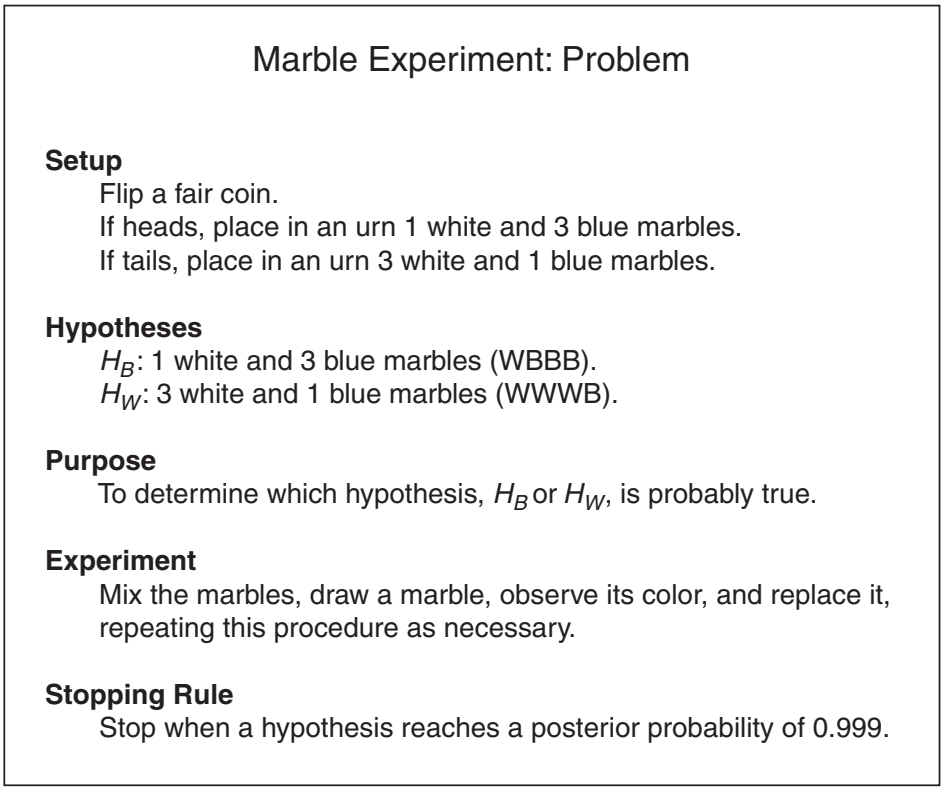

Figure 9.1 A marble experiment's setup, hypotheses, and purpose.

only 1 chance in 1,000 that the conclusion will be false. This marble problem is summarized in Figure 9.1.

The ratio form of Bayes's rule is convenient. Here it is recalled, with the earlier generic hypothesis labels "1" and "2" replaced by more informative labels, namely, "B" meaning mostly blue marbles (one white and three blue) and "W" meaning mostly white marbles (three white and one blue).

$$\frac{\mathrm{P}(H_B|E)}{\mathrm{P}(H_W|E)} = \frac{\mathrm{P}(E|H_B)}{\mathrm{P}(E|H_W)} \times \frac{\mathrm{P}(H_B)}{\mathrm{P}(H_W)} \tag{9.1}$$

Table 9.1 gives the data from an actual experiment with blue and white marbles and analyzes the data using this equation. From the coin toss, the prior odds for H_B:H_W are 1:1, so the prior probability $\mathrm{P}(H_B) = 0.5$, and this is also the posterior probability $\mathrm{P}(H_B \mid E) = 0.5$ before the experiment has generated any evidence.

The likelihood odds $\mathrm{P}(E \mid H_B)$:$\mathrm{P}(E \mid H_W)$ arising from each possible empirical outcome of drawing a blue or a white marble are as follows. Recalling that H_B has three of four marbles blue, but H_W has only one of four marbles blue, a blue draw is three times as probable given H_B as it is given H_W. Because $\mathrm{P}(\text{blue} \mid H_B) = 3/4 = 0.75$ and $\mathrm{P}(\text{blue} \mid H_W) = 1/4 = 0.25$, a blue draw contributes likelihood odds of 0.75:0.25 or 3:1 for H_B:H_W, favoring H_B. By similar reasoning, a white

Table 9.1 Bayesian analysis for an actual marble experiment, assuming prior odds for H_B:H_W of 1:1. The experiment concludes upon reaching a posterior probability of 0.999.

Draw	Outcome	Posterior H_B:H_W	Posterior $P(H_B \mid E)$
	(Prior)	1:1	0.500000
1	White	1:3	0.250000
2	Blue	1:1	0.500000
3	White	1:3	0.250000
4	Blue	1:1	0.500000
5	Blue	3:1	0.750000
6	Blue	9:1	0.900000
7	Blue	27:1	0.964286
8	Blue	81:1	0.987805
9	Blue	243:1	0.995902
10	White	81:1	0.987805
11	Blue	243:1	0.995902
12	White	81:1	0.987805
13	Blue	243:1	0.995902
14	Blue	729:1	0.998630
15	Blue	2187:1	0.999543

draw contributes likelihood odds of 1:3 against H_B. Furthermore, because each draw is an independent event after remixing the marbles, individual trials combine multiplicatively in an overall experiment. For example, two blue draws will generate likelihood odds in favor of H_B of 3:1 times 3:1, which equals 9:1. Thus, in a sequential experiment, each blue draw will increase the posterior odds for H_B:H_W by 3:1, whereas each white draw will decrease it by 1:3.

Applying this analysis to the data in Table 9.1, note that the first draw is a white marble, contributing likelihood odds of 1:3 against H_B. Multiplying those likelihood odds of 1:3 by the previous odds (the prior) of 1:1 gives posterior odds of 1:3, decreasing the posterior probability to $\mathrm{P}(H_B \mid E) = 0.25$, where the evidence at this point reflects one draw. In this sequential experiment, the posterior results after the first draw become the prior results at the start of the second draw. The second draw happens to be blue, contributing likelihood odds of 3:1 favoring H_B, thereby bringing the posterior probability $\mathrm{P}(H_B \mid E)$ back to the initial value of 0.5.

Moving on to the sixth draw, the previous odds are 3:1, and the current blue draw contributes likelihood odds of 3:1, resulting in posterior odds of 9:1 favoring H_B and hence a posterior probability of $\mathrm{P}(H_B \mid E) = 0.9$. Finally, after 15 draws, the posterior probability happens to exceed the stopping rule's

preselected value of 0.999, so the experiment stops, and hypothesis H_B is accepted with more than 99.9% probability of truth. Incidentally, in this particular instance of an actual marble experiment, the conclusion was indeed correct because the urn actually did contain three blue marbles and one white marble, as could have been demonstrated easily by some different experiment, such as drawing out all four marbles at once.

Table 9.1 illustrates an important feature of data analysis: results become more conclusive as an experiment becomes larger. During the first six draws, H_B has two wins, two losses, and two ties, so the results are quite inconclusive, and the better-supported hypothesis never reaches a probability beyond 0.9. Indeed, at only one draw and again at three draws, this experiment gives mild support to the false hypothesis! But draws 5 to 15 all give the win to H_B, which is actually true, finally with a probability greater than 0.999.

This particular experiment reached its verdict after 15 draws, but how long would such experiments be on average? A simple approximation, regardless whether H_B or H_W is true, is that on average each four draws give three draws that support the true hypothesis and one draw that supports the false hypothesis. Hence, on average, for four draws, two draws cancel out and two support the true hypothesis. Let M denote the margin of blue draws over white draws. Then, the posterior odds H_B:H_W equal 3^M:1, which exceed 999:1 or 99.9% confidence favoring H_B when $M = 7$, or exceed 1:999 favoring H_W when $M = -7$. Because half the data cancel and half count, the length L required for a margin of ± 7 averages about $2 \times 7 = 14$ draws. Hence, the particular experiment in Table 9.1, having 15 draws, is about average.

For M equal to 2, 3, 4, or 5, a more exact calculation gives the average length L as 3.2, 5.6, 7.8, or 9.9 draws, but thereafter the approximation that $L \approx 2M$ is quite accurate. For instance, if only 1 chance of error in 1,000,000 were to be tolerated, that would require a margin of 13 because $3^{13} = 1{,}594{,}323$ and hence an average length of about 26 draws. Because the weight of this experimental evidence grows exponentially with its amount, an exceedingly high probability of truth is readily attainable.

Furthermore, this exponential increase in the weight of the evidence confers robustness to this Bayesian analysis were this experiment to encounter various problems and complications that can plague real-world experiments. Problems can be disastrous but not necessarily so because weighty evidence can surmount considerable difficulties. Four substantial but surmountable problems are described here: controversial background information, messy data, wrong hypotheses, and different statistical methods.

Controversial Background Information. The foremost objection to Bayesian inference has been that frequently the background information that determines the prior probabilities is inadequate or even controversial. This perceived deficiency prompted the development of an alternative statistical paradigm, frequentist statistics, which will be described in the next section.

Recalling the mouse experiment in the previous chapter, which is analogous to the marble experiment in this chapter, Fisher judged that "the method of Bayes could properly be applied" because the pedigree information for these mice supplied "cogent knowledge" of the prior probabilities (Fisher 1973:8). On the other hand, "if knowledge of the origin of the mouse tested were lacking, no experimenter would feel he had warrant for arguing as if he knew that of which in fact he was ignorant, and for lack of adequate data" to determine the prior probabilities "Bayes' method of reasoning would be inapplicable" (Fisher 1973:20). Fisher's tale of the black and brown mice was a moral tale that waxed sermonic in its conclusion that "It is evidently easier for the practitioner of natural science to recognize the difference between knowing and not knowing than this seems to be for the more abstract mathematician," that is, for the Bayesian statistician (Fisher 1973:20).

For the present marble experiment, the prior probabilities $\mathrm{P}(H_B)$ and $\mathrm{P}(H_W)$ are known precisely because of the setup information about a coin toss. But what happens if no background information is given, so the prior probabilities are unknown and potentially controversial?

A particularly unfavorable case results from assigning a small prior probability to what is actually the true hypothesis, such as $\mathrm{P}(H_B) = 0.1$ when H_B is true. Prior odds for H_B:H_W of 1:9 require an additional likelihood odds of 9:1 to move the odds back to the 1:1 starting point of the original setup, which entails an average of about four draws. Hence, this unfavorable prior increases the original average length of the experiment from 14 to $14 + 4 = 18$ draws. Likewise, were the prior odds extremely challenging, such as $\mathrm{P}(H_B) = 0.001$ when H_B is true, the experimental effort increases to about 28 draws. Consequently, prior probabilities that are unfavorable to the truth result in more work, but the truth is still attainable.

A Bayesian statistician has essentially two alternatives for dealing with inadequate prior information. One alternative is to supply a noninformative prior, namely, $\mathrm{P}(H_B) = \mathrm{P}(H_W) = 0.5$, and also show what range of prior probabilities still leaves the conclusion unaltered, given the data at hand. If the data are strong, the conclusion may be robust despite a vague prior. The other alternative is to report the Bayes factor $\mathrm{P}(E \mid H_B) \,/\, \mathrm{P}(E \mid H_W)$ instead of the posterior probabilities because this avoids prior probabilities altogether. Either way, in the favorable case that the weight of the evidence grows exponentially with its amount, exceptionally strong evidence can be attainable that provides for a reliable and convincing conclusion.

Messy Data. In the original experimental procedure, the student, Beth, faithfully showed the class the marble resulting from each draw, ensuring quality data. But what happens if instead a weary and fickle Beth works alone and observes the drawn marble's color accurately only half of the time, whereas she reports blue or white at random the other half of the time? Can these messy data still decide between H_B and H_W with confidence?

The original margin between blue and white draws of seven draws allows the probability of a false conclusion to climb to 0.027 with the messy data. To maintain the specified 0.001 probability of error or 0.999 probability of truth, now the required margin increases to 14 and the average length of the experiment increases to about 56 draws. Hence, in this case, increased data quantity can compensate for decreased data quality. Of course, more pathological cases would be disastrous, such as unrecognized problems causing serious bias in the data. Frequently, scientific experiments are rather messy but not downright pathological, so the remedy of more data works.

Wrong Hypotheses. Certainly, one of the deepest problems that a scientific inquiry can possibly encounter is that the truth is not even among the hypotheses under consideration. For instance, consider this marble experiment with its setup specifying the hypotheses H_B with one white and three blue marbles, or else H_W with three white and one blue marbles. But what happens if by mistake or by mischief the experimentalist, Juan, puts two white and two blue marbles in the urn? Now the true hypothesis, H_E denoting equal numbers of both colors, is not even under consideration.

When H_E is true but only H_B and H_W are considered, on average, the experiment will require 49 draws until a margin of 7 draws declares H_B or else H_W true. But such a long experiment is suspicious, given that a length around 14 draws is expected. The length of the experiment has a wide variability around its average of 49 draws, with 70 or more draws occurring 22% of the time, which is extremely suspicious. But, on the other hand, only 20 or fewer draws occur 23% of the time, which would not be alarming. However, if the experiment were repeated several times, most likely the results would be weird: some unbelievably long experiments, contradictory conclusions favoring H_B about half of the time and H_W the other half, and frequencies for both blue and white draws near 0.5 for the pooled data. An unsuspected problem may escape detection after just one run but probably not after three or four runs, and almost certainly not after 10 or 20 runs.

The data are likely, at least eventually, to embarrass a faulty paradigm and thereby precipitate a paradigm shift. Even rather severe mistakes can be remediable. Scientific discovery is like a hike in the woods: you can go the wrong way for a while and yet still arrive at your destination at the end of the day.

Different Statistical Methods. Sometimes various scientists working on a given project adopt or prefer different statistical methods for various reasons, including debates between advocates of the Bayesian and frequentist approaches to statistics. How do statistical debates affect science? Can scientists get the same answers even if they apply different statistical methods to the data?

The short answer is that small experiments generating few data can leave scientists from different statistical schools with different conclusions about which hypothesis is most likely to be true. Rather frequently, scientists have only rather limited data, so the choice of a first-rate, efficient statistical procedure is

important. However, as more data become available, the influence of differences in statistical methods diminishes. Eventually, everyone will come to the same conclusion, even though they differ in terms of the particular calculations used and the exact confidence attributed to the unanimous conclusion.

Many additional challenges could be encountered beyond these four problems. For instance, closer hypotheses would be harder to discriminate between than H_B and H_W having widely separated probabilities of 0.75 and 0.25 for drawing a blue marble. The new hypotheses H_1 with three white and five blue marbles and H_2 with five white and three blue marbles give closer probabilities of 0.625 and 0.375. Now the average length of the experiments is about 56 draws to maintain the 0.999 probability of a true conclusion. Hence, data quantity can compensate for yet another potential challenge.

In conclusion, numerous problems can be overcome by the simple expedient of collecting more data, assuming that this option is not too expensive or difficult. This favorable outcome is especially likely when the weight of the evidence increases exponentially with the amount of the evidence.

Frequentist inference

Historically, the Bayesian paradigm preceded the frequentist paradigm by about a century and a half, so the latter was formulated in reaction to perceived problems with its predecessor. Principally, the frequentist paradigm sought to eliminate the Bayesian prior because it burdened scientists with the search for additional information that often was unavailable, diffuse, inaccurate, or controversial. Frequentists such as Sir Ronald A. Fisher, Jerzy Neyman, and Egon S. Pearson wanted to give scientists a paradigm with greater objectivity.

Frequentist statistics designates one hypothesis among those under consideration as the null hypothesis. Ordinarily, the null hypothesis is that there is no effect of the various treatments, whereas one or more alternative hypotheses express various possible treatment effects. A null hypothesis is either true or false, and a statistical test either accepts or rejects the null hypothesis, so there are four possibilities. A Type I error event is to reject a true null hypothesis, whereas a Type II error event is to accept a false null hypothesis.

The basic idea of frequentist hypothesis testing is that a statistical procedure with few Types I and II errors provides reliable learning from experiments. Type I errors can be avoided altogether merely by accepting every null hypothesis regardless of what the data show, and Type II errors can be avoided by rejecting every null. Hence, there is an inherent trade-off between Types I and II errors, so some compromise must be struck.

The ideal way to establish this compromise is to evaluate the cost or penalty for Type I errors and the cost for Type II errors and then balance those errors so as to minimize the overall expected cost of errors of both kinds. In routine

	True	False	
Accept	91	2	93
Reject	4	3	7
	95	5	100

Figure 9.2 Hypothetical example of error events and rates. A null hypothesis is either true or false, and a test, involving experimental data and a statistical inference, is used to accept or reject the null hypothesis, so there are four possible outcomes, with counts as shown. Also shown are row totals, column totals, and the grand total of 100 tests. To reject a true null hypothesis is a false-positive error event, whereas to accept a false null hypothesis is a false-negative error event.

practice, however, scientists tend to set the Type I error rate at some convenient level and not to be aware of the accompanying Type II error rate, let alone the implied overall or average cost of errors.

Figure 9.2 provides a concrete example. Such numbers could represent the results when a diagnostic test accepts or rejects a null hypothesis of no disease and, subsequently, a definitive test determines for sure whether the null is true or false.

Understand that an error *event* and an error *rate* are two different things. To reject a true null hypothesis is a Type I error event, and there are four such events. To accept a false null hypothesis is a Type II error event, and there are two such events. The Type I error rate α is $\mathrm{P}(\text{reject} \mid \text{true}) = 4/95 \approx 0.0421$ and the Type II error rate β is $\mathrm{P}(\text{accept} \mid \text{false}) = 2/5 = 0.4$. Note that the Type I error rate is $4/95$, not $4/7$ and not $4/100$.

Another important quantity for frequentists is the p-value, defined as the probability of getting an outcome at least as extreme as the actual observed outcome under the assumption that the null hypothesis is true. To calculate the p-value, one envisions repeating the experiment an infinite number of times and finds the probability of getting an outcome as extreme as or more extreme than the actual experimental outcome under the assumption that the null hypothesis is true. The smaller the p-value, the more strongly a frequentist test rejects the null hypothesis. It has become the convention in the scientific community to call rejection at a p-value of 0.05 a "significant" result and rejection at the 0.01 level a "highly significant" result.

To illustrate the calculation of a p-value, Table 9.2 analyzes the marble experiment from a frequentist perspective that was previously analyzed in Table 9.1 from a Bayesian perspective. Let the null hypothesis be H_W, that the urn contains three white marbles and one blue marble, and let the alternative hypothesis be

Table 9.2 Frequentist analysis for an actual marble experiment, assuming that the null hypothesis H_W is true and the experiment stops at 15 draws. At an experimental outcome of 11 blue draws (and 4 white draws), which is marked by an asterisk and is the same outcome as in Table 9.1, the conclusion is to reject H_W at the highly significant p-value of 0.000115.

Blue Draws	Probability	p-value
0	0.01336346101016	1.00000000000000
1	0.06681730505079	0.98663653898984
2	0.15590704511851	0.91981923393905
3	0.22519906517118	0.76391218882054
4	0.22519906517118	0.53871312364936
5	0.16514598112553	0.31351405847818
6	0.09174776729196	0.14836807735264
7	0.03932047169656	0.05662031006068
8	0.01310682389885	0.01729983836412
9	0.00339806545526	0.00419301446527
10	0.00067961309105	0.00079494901001
11	0.00010297168046	0.00011533591896 *
12	0.00001144129783	0.00001236423850
13	0.00000088009983	0.00000092294067
14	0.00000004190952	0.00000004284084
15	0.00000000093132	0.00000000093132

H_B, that it contains one white and three blue marbles. (In this case, neither hypothesis corresponds to the idea of no treatment effect, so H_W has been chosen arbitrarily to be the null hypothesis, but the story would be the same had H_B been designated the null hypothesis instead.)

Table 9.2 has three columns of numbers. The first column lists, for an experiment with 15 draws, the 16 possible outcomes, namely, 0 to 15 blue draws (and, correspondingly, 15 to 0 white draws). This analysis takes H_W as the null hypothesis, and under this assumption that the urn contains three white marbles and one blue marble, the probability of a blue draw is 0.25. So experiments with 15 draws will average $15 \times 0.25 = 3.75$ blue draws. Accordingly, were this experiment repeated many times, outcomes of about 3 or 4 blue draws would be expected to be rather frequent, whereas 14 or 15 blue draws would be quite rare. To upgrade this obvious intuition with an exact calculation using the probability theory explained in the preceding chapter, an outcome of b blue draws and w white draws from a total of $n = b + w$ draws can occur with $n! / (b! \times \mathrm{w}!)$ permutations, and the probability of each such outcome is

$0.25^b \times 0.75^w$. For example, the probability of 5 blue and 10 white draws is $[15! / (5! \times 10!)] \times 0.25^5 \times 0.75^{10} \approx 0.165146$.

These probabilities, for all possible outcomes from b values of 0 to 15, are listed in the second column of Table 9.2. Finally, the third column is the p-value, obtained for b blue draws by summing the probabilities for all outcomes with b or more blue draws. For example, the p-value for 0 blue draws is 1, because it is the sum of all 16 of these probabilities, whereas the p-value for 14 blue draws is the sum of the last two probabilities. For the particular marble experiment considered here, the actual outcome was 11 blue draws, and an asterisk draws attention to the corresponding p-value of 0.000115. The conclusion, based on this extremely small p-value, is to reject H_W as a highly significant result.

Unlike Bayesian analysis, which requires specification of prior probabilities in order to do the calculations, the frequentist analysis requires no such input, and thereby it seems admirably objective. So even if we know nothing about the process whereby the urn receives either the one white and three blue marbles or the reverse, we can still carry on unhindered with this wonderfully objective analysis! Or, so it seems.

Most persons who have read this section thus far probably have not sensed anything ambiguous or misleading in this frequentist analysis. It all seems so sensible. Besides, this statistical paradigm has dominated in scientific research for the previous several decades, so it hardly seems suspect. Nevertheless, there are some serious difficulties.

One problem is that, frequently, the error rate of primary concern to scientists is something other than the Type I or Type II error rates. The False Discovery Rate (FDR) is defined as the probability of the null hypothesis being true given that it is rejected, P(true | reject), which equals $4/7 \approx 0.5714$ for the example in Figure 9.2. It has the meaning here of the probability that a diagnosis of disease is actually false. Unfortunately, scientists often use the familiar Type I error rate P(reject | true) when their applications actually concern the FDR, which is the reverse conditional probability P(true | reject), which always has a different meaning and usually has a different value.

A worrisome feature of p-values is the strange influence accorded to the rule specifying when an experiment stops, which must be specified because every experiment must stop. The implicit stopping rule needed to make Table 9.2 comparable with Table 9.1 is that the experiment stops at 15 draws. But other stopping rules could result in exactly these same data, such as stop at 11 blue draws or stop at 4 white draws. However, these three rules differ in the imaginary outcomes that would result as the frequentist envisions numerous repetitions of the experiment. Consequently, exactly the same data can generate different p-values by assuming different stopping rules. For instance, Berger and Berry (1988) cited a disturbing example in which frequentist analyses of a single experiment gave p-values of 0.021, 0.049, 0.085, and any other value up to 1 just by assuming

different stopping rules. So a *p*-value depends on the experimental data *and* the stopping rule. Consequently, it depends on the actual experiment that did occur *and* an infinite number of other imaginary experiments that did not occur. Different stopping rules are generating different stories about just what those other imaginary experiments are, thereby changing *p*-values. But such reasoning seems bizarre and problematic, opening the door to unlimited subjectivity, quite in contradiction to the frequentists' grand quest for objectivity.

Another problem with *p*-values is that they usually overestimate, but can also underestimate, the strength of the evidence because they are strongly affected by the sample size. Raftery (1995) explained that Fisher's choice of Type I error rates α of 0.05 and 0.01 for significant and highly significant results were developed in the context of agricultural experiments with typical sample sizes in the range of 30 to 200, but these choices are misleading for sample sizes well outside this range. Contemporary experiments in the physical, biological, and social sciences often have sample sizes exceeding 10,000, for which the conventional $\alpha = 0.05$ will declare nearly all tests significant. For instance, from Raftery's Table 9, $\alpha = 0.053$ for a "significant" result with 50 samples corresponds to $\alpha = 0.0007$ with 100,000 samples, which is drastically different by a factor of almost 100. Unfortunately, many scientists are unaware of the adjustments in α that need to be made for sample size. Because of these problems with *p*-values, their use is declining, particularly in medical journals.

Bayesian methods have a tremendous advantage of computational ease over frequentist methods for models fitting thousands of parameters, which are becoming increasingly common in contemporary science. Also, some theorems (called complete class theorems) prove that even if one's objective is to optimize frequentist criteria, Bayesian procedures are often ideal for that (Robert 2007).

For introductory exposition of frequentist statistics, see Cox and Hinkley (1980); for Bayesian statistics, see Gelman et al. (2004) or Hoff (2009). For more technical presentations of Bayesian statistics, see the seminal text by Berger (1985) and the more recent text by Robert (2007).

Bayesian decision

The distinction between inference and decision is that inference problems pursue true beliefs, whereas decision problems pursue good actions. Clearly, inference and decision problems are interconnected because beliefs inform decisions and influence actions. Accordingly, decision problems incorporate inference sub-problems.

Many decisions are too simple or unimportant to warrant formal analysis, but some decisions are difficult and important. Formal decision analysis provides a logical framework that makes an individual's reasoning explicit, divides a complex problem into manageable components, eliminates inconsistencies in

a person's reasoning, clarifies the options, facilitates clear communication with others also involved in a decision, and promotes orderly and creative problem-solving. Sometimes life requires easy and quick decisions but at other times it demands difficult and careful decisions. Accordingly, formal decision methods are supplements to, not replacements for, informal methods. On the one hand, even modest study of formal decision theory can illuminate and refine ordinary informal decisions. On the other hand, simple common-sense decision procedures provide the only possible ultimate source and rational defense for a formal theory's foundations and axioms.

The basic structure of a decision problem is as follows. Decision theory partitions the components or causes of a situation into two fundamentally different groups on the basis of whether or not we have the power to control a given component or cause. What we can control is termed the "action" or choice. Obviously, to have a choice, there must exist at least two possible actions at our disposal. What we cannot control is termed the "state" or, to use a longer phrase, the state of nature. Each state-and-action combination is termed an "outcome," and each outcome is assigned a "utility" or "consequence" that assesses the value or benefit or goodness of that outcome, allowing negative values for loss or badness, and assigning zero for indifference. These possible consequences can be written in a consequences matrix, a two-way table with columns labeled with states and rows labeled with actions. There is also information on the probabilities of the states occurring, resulting from an inference sub-problem with its prior probabilities and likelihood information. If the state of nature were known or could be predicted with certainty, determining the best decision would be considerably easier; having only probabilistic information about the present or future state causes some complexity, uncertainty, and risk. Finally, the information on consequences and probabilities of states is combined in a decision criterion that assigns values to each choice and indicates the best action.

Figure 9.3 presents a simple example of a farmer's cropping decision. There are three possible states of nature, which are outside the farmer's control: good, fair, or bad weather. There are three possible actions among which the farmer can choose: plant crop A, plant crop B, or lease the land.

Beginning at the lower left portion of Figure 9.3, we know something about the probabilities of the weather states. We possess old and new data on the weather, summarized in the priors and likelihoods. For example, the old data could be long-run frequencies based on extensive historical climate records, indicating prior probabilities of 0.30, 0.50, and 0.20 for good, fair, and bad weather. The new data could be a recent long-range weather forecast that happens to favor good weather, giving likelihoods of 0.60, 0.30, and 0.10 for good, fair, and bad weather. Bayesian inference then combines the priors and likelihoods to derive the posterior probabilities of the weather states, as shown near the middle of the figure. Multiplying each prior by its corresponding

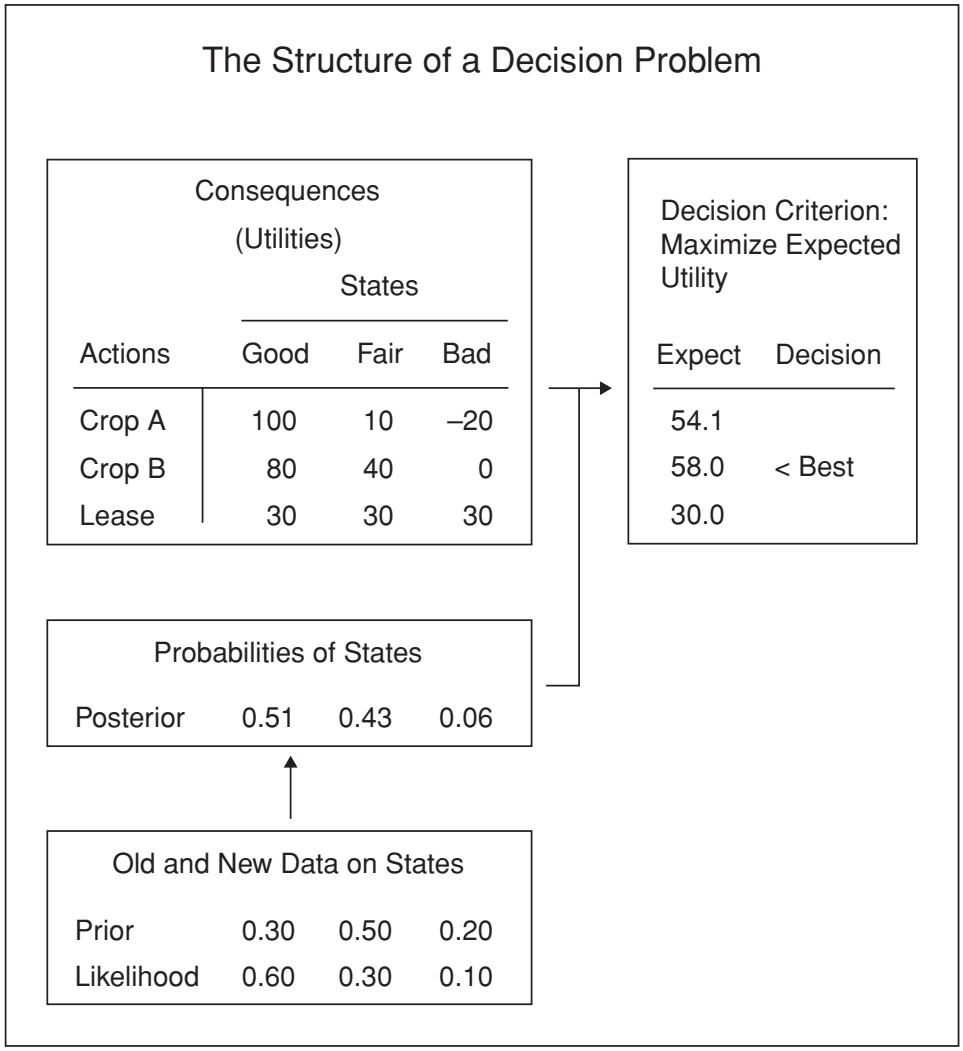

Figure 9.3 A decision problem about which crop to plant, which concludes that crop B is the best choice.

likelihood gives values of 0.18, 0.15, and 0.02, for a total of 0.35, and division of those three values by their total yields the posterior probabilities, namely, approximately 0.51, 0.43, and 0.06 for good, fair, and bad weather. So far, this is a standard inference problem. But a decision problem is more complicated, with two additional components, as explained next.

The upper left portion of Figure 9.3 shows the matrix of consequences or utilities. The outcome for any given growing season is specified by its particular state-and-action combination. The three possible states are good, fair, and bad weather, and the three possible actions are to plant crop A, plant crop B, or lease the land, for a total of $3 \times 3 = 9$ possible outcomes. The consequences matrix shows the utility or value of each possible outcome, using a positive number for a utility or gain, a negative number for a loss, or a zero for indifference. For example, in a given year, the outcome might be fair weather for crop B, which

has a utility of 40, where this number represents profit in dollars per acre or whatever.

Finally, the upper right portion of Figure 9.3 specifies a decision criterion, which is to maximize the expected utility. The expected utility is the average or predicted utility, calculated for each possible action by multiplying the utility for each state by its corresponding probability and summing over the states. For example, the expected utility for crop A is $(100 \times 0.51) + (10 \times 0.43) + (-20 \times 0.06) \approx 54.1$. Likewise, the expected utility for crop B is 58.0 and that for leasing is 30.0. The largest of these three values is 58.0, indicating that planting crop B is the best decision to maximize the expected utility.

This example illustrates a frequent feature of decision problems: different penalties for different errors can cause the best decision to differ from the best inference. Bayesian inference gives the greatest posterior probability of 0.51 to good weather, and good weather favors the choice of crop A. But Bayesian decision instead chooses crop B, with its largest expected utility of 58.0, primarily because fair weather is rather likely and would involve a tremendous reduction in crop A's utility.

Both probability and statistics require only the three Kolmogorov probability axioms (and the inherited predicate logic and arithmetic axioms), but decision theory requires the addition of one more axiom, such as the axiom of desirability of Jeffrey (1983:80–81). In essence, it says that the utility or desirability of an action equals the average of the utilities for its various outcomes weighted by their probabilities, as was done in Figure 9.3.

Because of different attitudes toward risk, decision criteria other than maximized expected utility may be appropriate and preferable. For example, one might prefer to minimize the worst possible utility, which in this case would favor leasing the land (because the worst possible utility from leasing would be 30, whereas crop A could be as bad as –20, and crop B as bad as 0). Sometimes the response to the expected utility is nonlinear, such as a strong response to utilities below some minimum needed for survival, but a mild response to differences among utilities that merely distinguish various levels of luxury. Furthermore, decisions can be evaluated in terms of not only their average but also their variability around that average, with large variability implying much uncertainty and risk. Sometimes a relatively minor compromise in the average can gain a substantial reduction in the variability, which is the basis for the insurance industry.

Decisions may have several criteria to be optimized simultaneously, probably with some complicated trade-offs and compromises. For example, a farmer might want to optimize income, as in Figure 9.3, but also want to rotate crops to avoid an epidemic buildup of pest populations and want to diversify crops to stagger the workload during busy seasons. Those other constraints might result in a decision, say, to plant 60% crop B and 40% crop A, which would reduce the expected utility slightly to $(0.6 \times 58.0) + (0.4 \times 54.1) \approx 56.4$.

Although decision problems are more complex than inference problems, in practice, they often are easier than inference problems because the necessity to take some action can allow even small probability differences to force sensible decisions. For example, other things being equal, even a slightly higher probability that a particular medicine is effective or a particular airplane is safe will suffice to generate strong preferences. So odds of merely 60:40 can force practical decisions. Because most probability reasoning is motivated by the practical need to make good decisions, not merely by theoretical interests, even rather weak data and small probability differences can still significantly inform and influence decisions.

Induction lost and regained

This chapter's account of inductive logic has been, on the whole, rather confident and cheerful. However, a tremendous philosophical battle has raged over induction from ancient Greek skeptics to the present, with David Hume's critique being especially well known. Without doubt, inductive logic has suffered more numerous and drastic criticisms than all of the other components of scientific reasoning combined. Dozens of books, mostly by philosophers, have been written on the so-called problem of induction.

Unfortunately, the verdict of history seems to be that "the salient feature of attempts to solve Hume's problem is that they have all failed" (Friedman 1990:28). Broad's oft-quoted aphorism says that induction is "the glory of science and the scandal of philosophy" (Broad 1952:143), and Whitehead (1925:25) called induction "the despair of philosophy." Howson (2000:14–15, 2) concluded that "Hume's argument is one of the most robust, if not the most robust, in the history of philosophy," and it simply is "actually correct."

Hume's critique of induction appeared in his anonymous, three-volume *A Treatise of Human Nature*, which was a commercial failure and drew heavy criticism from his fellow Scottish philosophers Thomas Reid and James Beattie. Subsequently, his admirably brief *An Enquiry Concerning Human Understanding* reformulated his critique, and that punchy book was a great success. Because Hume's advertisement in the latter work dismisses the former as a juvenile work, the discussion here follows the usual custom of examining just the *Enquiry*.

Hume's argument in Chapters 4 and 5 of his *Enquiry* has three key premises followed by the conclusion: (1) Any verdict on the legitimacy of induction must result from deductive or inductive arguments, because those are the only kinds of reasoning. (2) A verdict on induction cannot be reached deductively. No inference from the observed to the unobserved is deductive, specifically because nothing in deductive logic can ensure that the course of nature will not change. (3) A verdict cannot be reached inductively. Any appeal to the past successes of inductive logic, such as that bread has continued to be nutritious and that

the sun has continued to rise day after day, is but worthless circular reasoning when applied to induction's future fortunes. Therefore, because deduction and induction are the only options, and because neither can reach a verdict on induction, the conclusion follows that there is no rational justification for induction. Incidentally, whereas the second premise, that of no deductive link from the past to the future, had been well known since antiquity, the third premise, that of no (legitimate and noncircular) inductive link from the past to the future, was Hume's original and shocking innovation.

Induction suffered a second serious blow in the mid 1950s, two centuries after Hume, when Goodman (1955) propounded his "new riddle" of induction. "The new riddle of induction has become a well-known topic in contemporary analytic philosophy. . . . There are now something like twenty different approaches to the problem, or kinds of solutions, in the literature. . . . None of them has become the majority opinion, received answer, or textbook solution to the problem" (Douglas Stalker, in Stalker 1994:2).

Briefly, Goodman's argument ran as follows. Consider emeralds examined before time *t*, and suppose that all of them have been green (where *t* might be, say, tomorrow). The most simple and foundational inductive procedure, called the *straight rule of induction*, says that if a certain property has been found for a given proportion of many observed objects, then the same proportion applies to all similar unobserved objects as well as to individual unobserved objects. For example, if numerous rolls of a die have given an outcome of 2 with a frequency of nearly 1/6, then inductive logic leads us to the conclusion that the frequency of that outcome in all other rolls will also be 1/6, and likewise that the probability of any particular future roll giving that outcome will be 1/6. Similarly, those observations before time *t* of many emeralds that are all green support the inductive conclusion that all emeralds are green, as well as the prediction that if an emerald is examined after time *t*, it too will be green.

Then Goodman introduced a new property, "grue," with the definition that an object is grue if it is examined before time *t* and is green, or if it is not examined before time *t* and is blue. Admittedly, this is a rather contrived property, and the philosophical discussion of grue is quite technical and rather perplexing. But the main point is that Goodman showed that only some properties are appropriate (projectable) for applications of the straight rule of induction, but others are not. So how can one decide in a nonarbitrary manner which properties are projectable? Apart from clear criteria to discern when the straight rule is applicable, there is a danger that it will be used when inappropriate, thereby "proving" too much, even including contradictory conclusions.

All too predictably, Hume had complained that all received systems of philosophy were defective and impotent for justifying even the simple straight rule of induction. Goodman's complaint, however, was the exact opposite. His concern was not that induction proves too little but rather that it proves too much: a method that can prove anything proves nothing. Understand that

Goodman, like his predecessor Hume, was not intending to wean us from common sense, such as causing us to worry that all of our emeralds would turn from green to blue tomorrow. Rather, he was deploying the new riddle to wake us to the challenge of producing a philosophically respectable account of induction.

Finally, and perhaps most important, the great generality of those old and new problems of induction must be appreciated. Hume and Goodman expressed their arguments in terms of time: past and future, or before and after time *t*. But thoughtful commentators have discerned their broader scope. Gustason (1994:205) assimilated Hume's argument to a choice among various standard and nonstandard inductive logics. Accordingly, the resulting scope encompasses any and all inductive arguments, including those concerning exclusively past outcomes.

Howson (2000:30–32) followed Goodman in interpreting Goodman's argument as a demonstration that substantial prior knowledge about the world enters into our (generally sensible) choices about when to apply induction and how much data to require. Couvalis (1997:48) has cleverly said it all with a singularly apt example: "Having seen a large number of platypuses in zoos and none outside zoos, we do not infer that all platypuses live in zoos. However, having seen a small number of platypuses laying eggs, we might infer that all platypuses lay eggs." Similarly, Howson (2000:6, 197) observed that scientists are disposed to draw a sweeping generalization about the electrical conductivity of copper from measuring current flow in a few samples. But, obviously, many other scientific generalizations require enormous sample sizes.

Responding first to Hume's critique of induction, the role of common sense is critical. Hume said that we need not fear that doubts about induction "should ever undermine the reasonings of common life" because "Nature will always maintain her rights, and prevail in the end over any abstract reasoning whatsoever," and "Custom . . . is the great guide of human life" (Beauchamp 1999:120, 122). Hume's conclusion is not that induction is shaky but rather that induction is grounded in custom or habit or instinct, which we share with animals, rather than in philosophical reasoning. But Hume's argument depends on a controversial assumption that common sense is located outside philosophy rather than being an integral part and foundation of philosophy.

Indeed, when philosophy's roots in common sense are not honored, a characteristic pathology ensues: instead of natural philosophy happily installing science's presuppositions once, at the outset, by faith in a trifling trinket of common-sense knowledge, a death struggle with skepticism gets repeated over and over again for each component of scientific method, including induction. The proper task, "to explain induction," swells to the impossible task, "to defeat skepticism and explain induction." If Hume's philosophy cannot speak in induction's favor, that is because it is a truncated version of philosophy that has exiled animal habit rather than having accommodated our incarnate human

nature as an integral component of philosophy's common-sense starting points, as Reid had recommended.

Plainly, all of the action in Hume's attack on induction derives ultimately from the concern that the course of nature might change, but that is simply the entrance of skepticism. His own examples include such drastic matters as whether or not the sun will continue to rise daily and bread will continue to be nutritious. Such matters are nothing less than philosophy's ancient death fight with skepticism! They are nothing less than the end of the world! In the apocalypse proposed by those examples, not only does induction hang in the balance but also planetary orbits and biological life. As Himsworth (1986:87–88) observed in his critique of Hume, if the course of nature did change, we would not be here to complain! So as long as we are here or we are talking about induction, deep worries about induction are unwarranted. Consequently, seeing that apocalypse as "the problem of induction" rather than "the end of the world" is like naming a play for an incidental character. The rhetoric trades in obsessive attention to one detail.

Turning next to Goodman's new riddle of induction, it shows that although the straight rule of induction is itself quite simple, judging whether or not to apply it to a given property for a given sample is rather complicated. These judgments, as in the example of platypuses, draw on general knowledge of the world and common sense. Such broad and diffuse knowledge resists tidy philosophical analysis.

Summary

Induction reasons from actual data to an inferred model, whereas deduction reasons from a given model to expected data. Both are important for science, composing the logic or "L" portion of the PEL model. Probability is the deductive science of uncertainty, whereas statistics is the inductive science of uncertainty.

Aristotle, medieval philosopher-scientists, and modern scholars have developed various inductive methods. But not until the publication in 1763 of Bayes's theorem was the problem finally solved of relating conditional probabilities of the form $P(E \mid H)$, the probability of evidence E given hypothesis H, found in deduction with reverse conditional probabilities of the form $P(H \mid E)$ required in induction. Bayes's theorem was illustrated with a simple example regarding blue and white marbles drawn from an urn. Inductive conclusions can be robust despite considerable difficulties with controversial background information, messy data, wrong hypotheses, and different statistical methods. Particularly when the weight of the evidence grows exponentially with the amount of the evidence, increased data quantity can often compensate for decreased data quality. However, the need to specify prior probabilities, which can be unknown

or even controversial, prompted the development of an alternative paradigm intended to be more applicable and objective.

Frequentist inference, which is a competitor to Bayesian inference, was illustrated with the same marble experiment. Frequentist methods seek to minimize Type I errors, rejecting a true null hypothesis, and Type II errors, accepting a false null hypothesis, but this is challenging because of the inevitable trade-off between these two kinds of errors. Statistical significance is assessed by *p*-values that express the probability of getting an outcome as extreme, or more extreme, than the actual experimental outcome under the assumption that the null hypothesis is true. But sometimes error rates other than the Type I and Type II error rates are more relevant, particularly the False Discovery Rate. And *p*-values have been criticized because their strange dependence on stopping rules and imaginary outcomes undermines the presumed pursuit of objectivity and because their actual significance depends strongly on the number of samples.

Bayesian decision theory was illustrated with a simple example of a farmer's cropping decision. Whereas inference problems pursue true beliefs, decision problems pursue good actions. Decision theory requires one more axiom beyond those already needed for probability and statistics, an axiom of desirability saying in essence that the utility or desirability of an action equals the average of the utilities for its various outcomes weighted by their probabilities.

Inductive logic has received far more philosophical criticism than all of the other components of scientific method combined. David Hume argued that philosophy cannot justify any inductive procedures, including the simple straight rule of induction. More recently, Goodman's new riddle of induction showed the exact opposite, that the straight rule of induction can be used to prove anything – which is equally problematic. But given common-sense presuppositions, induction can be defended and implemented effectively.

Study questions

(1) Recall that the vertical bar in a conditional probability is read as "given," so $P(A \mid B)$ means the probability of A given B. Let H denote a hypothesis and E denote some evidence. How do $P(H \mid E)$ and $P(E \mid H)$ differ in meaning and in numerical value? How are they related by Bayes's theorem? How do they pertain to scientists' main research questions?

(2) List several kinds of problems that sometimes plague scientific experiments. How can inductive logic or statistics reach reliable and robust conclusions despite such problems?

(3) Define and compare Type I, Type II, and False Discovery Rate (FDR) error rates. Suppose that your research has two steps: an inexpensive initial screening for numerous promising candidates, followed by a very expensive

final test for promising candidates. Which kind of error rate would be most relevant for the initial screening and why?

(4) Does either the Bayesian or frequentist paradigm have a legitimate claim overall to greater objectivity and, if so, for exactly what reasons? What is the relative importance of statistical paradigm and evidential strength in achieving objectivity?

(5) Describe Hume and Goodman's riddles of induction. What are your own responses to these riddles? Do they undermine induction or not? What role do common-sense presuppositions play in a defense of induction?

References

AAAS (American Association for the Advancement of Science). 1989. *Science for All Americans: A Project 2061 Report on Literacy Goals in Science, Mathematics, and Technology*. Washington, DC: AAAS.

1990. *The Liberal Art of Science: Agenda for Action*. Washington, DC: AAAS.

1993. *Benchmarks for Science Literacy*. Oxford, UK: Oxford University Press.

2000. *Designs for Science Literacy*. Oxford, UK: Oxford University Press.

Abd-El-Khalick, F., and Lederman, N. G. 2000. Improving science teachers' conceptions of nature of science: A critical review of the literature. *International Journal of Science Education* 22:665–701.

Adler, M. J. 1978. *Aristotle for Everybody: Difficult Thought Made Easy*. New York: Macmillan.

Akerson, V. L., Buck, G. A., Donnelly, L. A., Nargund-Joshi, V., and Weiland, I. S. 2011. The importance of teaching and learning nature of science in the early childhood years. *Journal of Science Education and Technology* 20: 537–549.

Annas, J., and Barnes, J. 1985. *The Modes of Scepticism: Ancient Texts and Modern Interpretations*. Cambridge, UK: Cambridge University Press.

Anscombe, G. E. M., and von Wright, G. H. (eds.). 1969. *On Certainty* (by Ludwig Wittgenstein). New York: Harper & Row.

Ark, T. K., Brooks, L. R., and Eva, K. W. 2007. The benefits of flexibility: The pedagogical value of instructions to adopt multifaceted diagnostic reasoning strategies. *Medical Education* 41:281–287.

Audi, R. (ed.). 1999. *The Cambridge Dictionary of Philosophy*, 2nd edn. Cambridge, UK: Cambridge University Press.

Banner, M. C. 1990. *The Justification of Science and the Rationality of Religious Belief*. Oxford, UK: Oxford University Press.

Barnard, G. A. 1958. Studies in the history of probability and statistics. IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances. *Biometrika* 45:293–315.

Bautista, N. U., and Schussler, E. E. 2010. Implementation of an explicit and reflective pedagogy in introductory biology laboratories. *Journal of College Science Teaching* 40:56–61.

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 53:370–418.
Beauchamp, T. L. (ed.). 1999. *David Hume: An Enquiry Concerning Human Understanding*. Oxford, UK: Oxford University Press.
Becker, B. J. 2000. MindWorks: Making scientific concepts come alive. *Science & Education* 9:269–278.
Ben-Chaim, D., Ron, S., and Zoller, U. 2000. The disposition of eleventh-grade science students toward critical thinking. *Journal of Science Education and Technology* 9:149–159.
Benjamin, A. S., and Hackstaff, L. H. 1964. *Saint Augustine: On Free Choice of the Will*. Indianapolis, IN: Bobbs-Merrill.
Berger, J. O. 1985. *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. New York: Springer-Verlag.
Berger, J. O., and Berry, D. A. 1988. Statistical analysis and the illusion of objectivity. *American Scientist* 76:159–165.
Bering, J. M. 2006. The cognitive psychology of belief in the supernatural. *American Scientist* 94:142–149.
Berry, D. A. 1987. Interim analysis in clinical trials: The role of the likelihood principle. *American Statistician* 41:117–122.
Blackburn, S. 1994. *The Oxford Dictionary of Philosophy*. Oxford, UK: Oxford University Press.
Boardman, J. D., Blalock, C. L., Corley, R. P., et al. 2010. Ethnicity, body mass, and genome-wide data. *Biodemography and Social Biology* 56:123–136.
Boehner, P. 1957. *Ockham Philosophical Writings*. New York: Nelson.
Boghossian, P. 2006. *Fear of Knowledge: Against Relativism and Constructivism*. Oxford, UK: Oxford University Press.
Boudreau, C., and McCubbins, M. D. 2009. Competition in the courtroom: When does expert testimony improve jurors' decisions? *Journal of Empirical Legal Studies* 6:793–817.
Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. New York: John Wiley & Sons, Ltd.
Boyd, R., Gasper, P., and Trout, J. D. (eds.). 1991. *The Philosophy of Science*. Cambridge, MA: MIT Press.
Broad, C. D. 1952. *Ethics and the History of Philosophy*. London: Routledge & Kegan Paul.
Broad, W. J. 1979. Paul Feyerabend: Science and the anarchist. *Science* 206:534–537.
Brown, H. I. 1987. *Observation and Objectivity*. Oxford, UK: Oxford University Press.
Brown, S. C. (ed.). 1984. *Objectivity and Cultural Divergence*. Cambridge, UK: Cambridge University Press.
Brünger, A. T. 1992. Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355:472–475.
1993. Assessment of phase accuracy by cross validation: The free R value. *Acta Crystallographica, Section D* 49:24–36.
Buera, F. J., Monge-Naranjo, A., and Primiceri, G. E. 2011. Learning the wealth of nations. *Econometrica* 79:1–45.

Burks, A. W. 1977. *Chance, Cause, Reason: An Inquiry into the Nature of Scientific Evidence*. Chicago, IL: University of Chicago Press.

Cajori, F. 1947. *Sir Isaac Newton's Mathematical Principles of Natural Philosophy and His System of the World*. Berkeley, CA: University of California Press.

Caldin, E. F. 1949. *The Power and Limits of Science: A Philosophical Study*. London: Chapman & Hall.

Callebaut, W. 1993. *Taking the Naturalistic Turn, or How Real Philosophy of Science Is Done*. Chicago, IL: University of Chicago Press.

Callus, D. A. (ed.). 1955. *Robert Grosseteste, Scholar and Bishop: Essays in Commemoration of the Seventh Centenary of His Death*. Oxford, UK: Oxford University Press.

Cameron, C. M., and Park, J.-K. 2009. How will they vote? Predicting the future behavior of Supreme Court nominees, 1937–2006. *Journal of Empirical Legal Studies* 6:485–511.

Carey, S. S. 2012. *A Beginner's Guide to Scientific Method*, 4th edn. Belmont, CA: Wadsworth.

Cartier, J. L., and Stewart, J. 2000. Teaching the nature of inquiry: Further developments in a high school genetics curriculum. *Science & Education* 9:247–267.

Carus, P. 1902. *Immanuel Kant: Prolegomena to Any Future Metaphysics That Can Qualify as a Science*. La Salle, IL: Open Court.

Chatalian, G. 1991. *Epistemology and Skepticism: An Enquiry into the Nature of Epistemology*. Carbondale, IL: Southern Illinois University Press.

Clark, J. S. 2005. Why environmental scientists are becoming Bayesians. *Ecology Letters* 8:2–14.

Cleland, C. E. 2011. Prediction and explanation in historical natural science. *The British Journal for the Philosophy of Science* 62:551–582.

Cobern, W. W., and Loving, C. C. 2001. Defining "science" in a multicultural world: Implications for science education. *Science Education* 85:50–67.

Collins, F. S. 2006. *The Language of God: A Scientist Presents Evidence for Belief*. New York: Free Press.

Collins, H. 2009. We cannot live by scepticism alone. *Nature* 458:30–31, correspondence 458:702–703.

Collins, H., and Pinch, T. 1993. *The Golem: What Everyone Should Know About Science*. Cambridge, UK: Cambridge University Press.

Couvalis, G. 1997. *The Philosophy of Science: Science and Objectivity*. London: Sage.

Cox, D. R., and Hinkley, D. V. 1980. *Problems and Solutions in Theoretical Statistics*. New York: Chapman & Hall.

Craig, W. L., and Moreland, J. P. (eds.). 2009. *The Blackwell Companion to Natural Theology*. Chichester, UK: Wiley-Blackwell.

Crombie, A. C. 1962. *Robert Grosseteste and the Origins of Experimental Science*. Oxford, UK: Oxford University Press.

Cullen, T. A., Akerson, V. L., and Hanson, D. L. 2010. Using action research to engage K–6 teachers in nature of science inquiry as professional development. *Journal of Science Teacher Education* 21:971–992.

Cuneo, T., and van Woudenberg, R. (eds.). 2004. *The Cambridge Companion to Thomas Reid*. Cambridge, UK: Cambridge University Press.

Dales, R. C. 1973. *The Scientific Achievement of the Middle Ages*. Philadelphia: University of Pennsylvania Press.

Dalgarno, M., and Matthews, E. (eds.). 1989. *The Philosophy of Thomas Reid*. Dordrecht: Kluwer.
Dampier, W. C. 1961. *A History of Science and Its Relations with Philosophy and Religion*. Cambridge, UK: Cambridge University Press.
Davies, B., and Leftow, B. (eds.). 2006. *Thomas Aquinas: Summa Theologiae, Questions on God*. Cambridge, UK: Cambridge University Press.
Dawkins, R. 1996. *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe Without Design*. New York: Norton.
2006. *The God Delusion*. Boston, MA: Houghton Mifflin.
Derry, G. N. 1999. *What Science is and How it Works*. Princeton, NJ: Princeton University Press.
Dixon, T., Cantor, G., and Pumfrey, S. (eds.). 2010. *Science and Religion: New Historical Perspectives*. Cambridge, UK: Cambridge University Press.
Douglas, H. E. 2009. Reintroducing prediction to explanation. *Philosophy of Science* 76:444–463.
Earman, J. 2000. *Hume's Abject Failure: The Argument against Miracles*. Oxford, UK: Oxford University Press.
Easterbrook, G. 1997. Science and God: A warming trend? *Science* 277:890–893.
Ecklund, E. H. 2010. *Science vs. Religion: What Scientists Really Think*. Oxford, UK: Oxford University Press.
Edwards, P. (ed.). 1967. *The Encyclopedia of Philosophy*, 8 vols. New York: Macmillan.
Eggen, P.-O., Kvittingen, L., Lykknes, A., and Wittje, R. 2012. Reconstructing iconic experiments in electrochemistry: Experiences from a history of science course. *Science & Education* 21:179–189.
Ehrlich, I. 1973. Participation in illegitimate activities: A theoretical and empirical investigation. *Journal of Political Economy* 81:521–565.
Einstein, A. 1954. *Ideas and Opinions*. New York: Crown.
Eisenberg, T., and Wells, M. T. 2008. Statins and adverse cardiovascular events in moderate-risk females: A statistical and legal analysis with implications for FDA preemption claims. *Journal of Empirical Legal Studies* 5:507–550.
Finkelstein, M. O. 2009. *Basic Concepts of Probability and Statistics in the Law*. New York: Springer.
Fisher, R. A. 1973. *Statistical Methods and Scientific Inference*, 3rd edn. New York: Macmillan. (Originally published 1956.)
Frank, P. G. (ed.). 1957. *Philosophy of Science: The Link Between Science and Philosophy*. Englewood Cliffs, NJ: Prentice-Hall.
Frege, G. 1893. *Grundgesetze der Arithmetik*. Jena: Hermann Pohle.
Freireich, E. J., Gehan, E., Frei, E., et al. 1963. The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. *Blood* 21:699–716.
Friedman, K. S. 1990. *Predictive Simplicity: Induction Exhum'd*. Elmsford Park, NY: Pergamon.
Gauch, H. G. 1988. Model selection and validation for yield trials with interaction. *Biometrics* 44:705–715.
1992. *Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs*. New York: Elsevier.

1993. Prediction, parsimony, and noise. *American Scientist* 81:468–478, correspondence 507–508.

2002. *Scientific Method in Practice*. Cambridge, UK: Cambridge University Press; Chinese edition, 2004, Beijing: Tsinghua University Press.

2006. Winning the accuracy game. *American Scientist* 94:133–141, correspondence 94:196, addendum 94:382.

2009a. Science, worldviews, and education. *Science & Education* 18:667–695; also published in Matthews, M. R. (ed.). 2009. *Science, Worldviews and Education*, pp. 27–48. Heidelberg: Springer Verlag.

2009b. Responses and clarifications regarding science and worldviews. *Science & Education* 18:905–927; also published in Matthews, M. R. (ed.). 2009. *Science, Worldviews and Education*, pp. 303–325. Heidelberg: Springer Verlag.

Gauch, H. G., Rodrigues, P. C., Munkvold, J. D., Heffner, E. L., and Sorrells, M. 2011. Two new strategies for detecting and understanding QTL × environment interactions. *Crop Science* 51:96–113.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. 2004. *Bayesian Data Analysis*, 2nd edn. New York: Chapman & Hall.

Gentner, D., Loewenstein, J., and Thompson, L. 2003. Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology* 95:393–408.

Giacomini, R., and White, H. 2006. Tests of conditional predictive ability. *Econometrica* 74:1545–1578.

Gimbel, S. (ed.). 2011. *Exploring the Scientific Method: Cases and Questions*. Chicago, IL: University of Chicago Press.

Glynn, L. 2011. A probabilistic analysis of causation. *British Journal for the Philosophy of Science* 62:343–392.

Godfrey-Smith, P. 2003. *Theory and Reality: An Introduction to the Philosophy of Science*. Chicago, IL: University of Chicago Press.

Good, R. P., Kost, D., and Cherry, G. A. 2010. Introducing a unified PCA algorithm for model size reduction. *IEEE Transactions on Semiconductor Manufacturing* 23:201–209.

Goodman, L. A., and Hout, M. 2002. Statistical methods and graphical displays for analyzing how the association between two qualitative variables differs among countries, among groups, or over time: A modified regression-type approach. *Statistical Methodology* 28:175–230.

Goodman, N. 1955. *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Gottfried, K., and Wilson, K. G. 1997. Science as a cultural construct. *Nature* 386:545–547.

Gower, B. 1997. *Scientific Method: An Historical and Philosophical Introduction*. London: Routledge.

Grayling, A. C. 2011. How we form beliefs. *Nature* 474:446–447.

Greenland, S. 2008. Introduction to Bayesian statistics. In: Rothman, K. J., Greenland, S., and Lash, T. L. (eds). *Modern Epidemiology*, 3rd ed., pp. 328–344. New York: Wolters Kluwer.

Grier, B. 1975. Prediction, explanation, and testability as criteria for judging statistical theories. *Philosophy of Science* 42:373–383.

Gross, P. R., Levitt, N., and Lewis, M. W. (eds.). 1996. *The Flight from Science and Reason*. Baltimore, MD: Johns Hopkins University Press.
Gustason, W. 1994. *Reasoning from Evidence: Inductive Logic*. New York: Macmillan.
Guyer, P. (ed.). 1992. *The Cambridge Companion to Kant*. Cambridge, UK: Cambridge University Press.
Hamilton, A. G. 1978. *Logic for Mathematicians*. Cambridge, UK: Cambridge University Press.
Hamilton, W. (ed.). 1872. *The Works of Thomas Reid, D.D.*, 7th edn. Edinburgh: MacLachlan & Stewart.
Himsworth, H. 1986. *Scientific Knowledge and Philosophic Thought*. Baltimore, MD: Johns Hopkins University Press.
Hodson, D. 2009. *Teaching and Learning about Science: Language, Theories, Methods, History, Traditions and Values*. Rotterdam: Sense Publishers.
Hoff, P. D. 2009. *A First Course in Bayesian Statistical Methods*. New York: Springer.
Hoffmann, R., Minkin, V. I., and Carpenter, B. K. 1996. Ockham's razor and chemistry. *Bulletin de la Société chimique de France* 133(2):117–130.
Horgan, J. 1991. Profile: Thomas S. Kuhn, reluctant revolutionary. *Scientific American* 264(5):40, 49.
1992. Profile: Karl R. Popper, the intellectual warrior. *Scientific American* 267(5): 38–44.
1993. Profile: Paul Karl Feyerabend, the worst enemy of science. *Scientific American* 268(5):36–37.
2005. Clash in Cambridge: Science and religion seem as antagonistic as ever. *Scientific American* 293(3):24B, 26–28.
Howson, C. 2000. *Hume's Problem: Induction and the Justification of Belief*. Oxford, UK: Oxford University Press.
Howson, C., and Urbach, P. 2006. *Scientific Reasoning: The Bayesian Approach*, 3rd edn. La Salle, IL: Open Court.
Hunter, A.-B., Laursen, S. L., and Seymour, E. 2007. Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development. *Science Education* 91:36–74.
Irwin, T. H. 1988. *Aristotle's First Principles*. Oxford, UK: Oxford University Press.
Irzik, G., and Nola, R. 2010. A family resemblance approach to the nature of science for science education. *Science & Education* 20:591–607.
Jefferys, W. H., and Berger, J. O. 1992. Ockham's razor and Bayesian analysis. *American Scientist* 80:64–72.
Jeffrey, R. C. 1983. *The Logic of Decision*, 2nd edn. Chicago, IL: University of Chicago Press.
Jeffreys, H. 1973. *Scientific Inference*, 3rd edn. Cambridge, UK: Cambridge University Press.
1983. *Theory of Probability*, 3rd edn. Oxford, UK: Oxford University Press.
Johnson, M. A., and Lawson, A. E. 1998. What are the relative effects of reasoning ability and prior knowledge on biology achievement in expository and inquiry classes? *Journal of Research in Science Teaching* 35:89–103.
Joo, Y., Wells, M. T., and Casella, G. 2010. Model selection error rates in nonparametric and parametric model comparisons. In Berger, J. O., Cai, T. T., and Johnstone, I. M.

(eds.). *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown*, pp. 166–183. Beachwood, OH: Institute of Mathematical Statistics.

Kass, R. E., and Raftery, A. E. 1995. Bayes factors. *Journal of the American Statistical Association* 90:773–795.

Kemeny, J. G. 1959. *A Philosopher Looks at Science*. New York: Van Nostrand.

Khatib, F., DiMaio, F., Foldit Contenders Group, Foldit Void Crushers Group, et al. 2011. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature Structural & Molecular Biology* 18:1175–1177.

Khine, M. S. (ed.). 2011. *Advances in the Nature of Science Research: Concepts and Methodologies*. New York: Springer.

King, P. 1995. *Augustine: Against the Academicians; The Teacher*. Indianapolis, IN: Hackett.

Kleywegt, G. J. 2007. Separating model optimization and model validation in statistical cross-validation as applied to crystallography. *Acta Crystallographica Section D* 63:939–940.

Kleywegt, G. J., and Jones, T. A. 1995. Where freedom is given, liberties are taken. *Structure* 3:535–540.

2002. Homo Crystallographicus – Quo Vadis? *Structure* 10:465–472.

Koertge, N. (ed.). 1998. *A House Built on Sand: Exposing Postmodernist Myths about Science*. Oxford, UK: Oxford University Press.

Kolmogorov, A. N. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Julius Springer.

Kuhn, T. S. 1970. *The Structure of Scientific Revolutions*, 2nd edn. Chicago, IL: University of Chicago Press. (Originally published 1962.)

Lakatos, I., and Musgrave, A. (eds.). 1968. *Problems in the Philosophy of Science*. Amsterdam: North-Holland.

(eds.). 1970. *Criticism and the Growth of Knowledge*. Cambridge, UK: Cambridge University Press.

Laksov, K. B., Lonka, K., and Josephson, A. 2008. How do medical teachers address the problem of transfer? *Advances in Health Sciences Education* 13: 345–360.

Larson, E. J., and Witham, L. 1999. Scientists and religion in America. *Scientific American* 281(3):88–93.

Lawson, A. E., Clark, B., Cramer-Meldrum, E., Falconer, K. A., Sequist, J. M., and Kwon, Y.-J. 2000. Development of scientific reasoning in college biology: Do two levels of general hypothesis-testing skills exist? *Journal of Research in Science Teaching* 37:81–101.

Lawson, A. E., and Weser, J. 1990. The rejection of nonscientific beliefs about life: Effects of instruction and reasoning skills. *Journal of Research in Science Teaching* 27:589–606.

Lederman, N. G. 1992. Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching* 29:331–359.

2007. Nature of Science: Past, Present, and Future. In: Abell, S. K., and Lederman, N. G. (eds.). *Handbook of Research on Science Education*, pp. 831–879. Mahwah, NJ: Lawrence Erlbaum Associates.

Lehrer, K. 1989. *Thomas Reid*. London: Routledge.

Leitgeb, H., and Pettigrew, R. 2010a. An objective justification of Bayesianism. I: Measuring inaccuracy. *Philosophy of Science* 77:201–235.

2010b. An objective justification of Bayesianism. II: The consequences of minimizing inaccuracy. *Philosophy of Science* 77:236–272.

Leplin, J. 1997. *A Novel Defense of Scientific Realism*. Oxford, UK: Oxford University Press.

Lide, D. R. (ed.). 1995. *CRC Handbook of Chemistry and Physics*, 76th edn. Boca Raton, FL: CRC Press.

Lindberg, D. C. 2007. *The Beginnings of Western Science: The European Scientific Tradition in Philosophical, Religious, and Institutional Context, Prehistory to A.D. 1450*, 2nd edn. Chicago, IL: University of Chicago Press.

Lindberg, D. C., and Numbers, R. L. 2003. *When Science and Christianity Meet*. Chicago, IL: University of Chicago Press.

Lloyd, E. A. 2010. Confirmation and robustness of climate models. *Philosophy of Science* 77:971–984.

Lobato, J. 2006. Alternative perspectives on the transfer of learning: History, issues, and challenges for future research. *The Journal of the Learning Sciences* 15: 431–449.

Losee, J. 2001. *A Historical Introduction to the Philosophy of Science*, 4th edn. Oxford, UK: Oxford University Press.

2011. *Theories of Causality: From Antiquity to the Present*. New Brunswick, NJ: Transaction Publishers.

Lv, J. C., Yi, Z., and Tan, K. K. 2007. Determination of the number of principal directions in a biologically plausible PCA model. *IEEE Transactions on Neural Networks* 18:910–916.

MacDonald, S. 1998. Natural theology. In: Craig, E. (ed.). *Routledge Encyclopedia of Philosophy*, vol. 6, pp. 707–713. London: Routledge.

MacIntyre, A. 1988. *Whose Justice? Which Rationality?* Notre Dame, IN: University of Notre Dame Press.

MacKay, D. J. C. 1992. Bayesian interpolation. *Neural Computation* 4:415–447.

Marrone, S. P. 1983. *William of Auvergne and Robert Grosseteste: New Ideas of Truth in the Early Thirteenth Century*. Princeton, NJ: Princeton University Press.

Matthews, M. R. 1994. *Science Teaching: The Role of History and Philosophy of Science*. London: Routledge.

1998. In defense of modest goals when teaching about the nature of science. *Journal of Research in Science Teaching* 35:161–174.

2000. *Time for Science Education: How Teaching the History and Philosophy of Pendulum Motion Can Contribute to Science Literacy*. Dordrecht: Kluwer.

McClellan, J. E., and Dorn, H. 2006. *Science and Technology in World History: An Introduction*, 2nd edn. Baltimore, MD: Johns Hopkins University Press.

McCloskey, M. 1983. Intuitive physics. *Scientific American* 248(4):122–130.

McComas, W. F. (ed.). 1998. *The Nature of Science in Science Education: Rationales and Strategies*. Dordrecht: Kluwer.

McGrayne, S. B. 2011. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines & Emerged Triumphant from Two Centuries of Controversy*. New Haven, CT: Yale University Press.

McGrew, T. 2003. Confirmation, heuristics, and explanatory reasoning. *The British Journal for the Philosophy of Science* 54:553–567.
McKeon, R. 1941. *The Basic Works of Aristotle*. New York: Random House.
McQuarrie, A. D. R., and Tsai, C.-L. 1998. *Regression and Time Series Model Selection*. River Edge, NJ: World Scientific.
Medawar, P. B. 1969. *Induction and Intuition in Scientific Thought*. Philadelphia: American Philosophical Society.
Merton, R. K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago, IL: University of Chicago Press.
Merton, R. K., and Sztompka, P. 1996. *On Social Structure and Science*. Chicago, IL: University of Chicago Press.
Meyling, H. 1997. How to change students' conceptions of the epistemology of science. *Science & Education* 6:397–416.
Moore, J. H. 1998. Public understanding of science – and other fields. *American Scientist* 86:498.
Mukerjee, M. 1998. Undressing the emperor. *Scientific American* 278(3):30, 32.
Myrvold, W. C. 1996. Bayesianism and diverse evidence. *Philosophy of Science* 63:661–665.
2003. A Bayesian account of the virtue of unification. *Philosophy of Science* 70: 399–423.
NAS (National Academy of Sciences). 1995. *Reshaping the Graduate Education of Scientists and Engineers*. Washington, DC: National Academies Press.
NAS (and National Academy of Engineering and Institute of Medicine). 2009. *On Being a Scientist: A Guide to Responsible Conduct in Research*, 3rd edn. Washington, DC: National Academies Press.
NAS. 2010. *Managing University Intellectual Property in the Public Interest*. Washington, DC: National Academies Press.
Nash, L. K. 1963. *The Nature of the Natural Sciences*. Boston, MA: Little, Brown.
NCEE (National Commission on Excellence in Education). 1983. *A Nation at Risk: The Imperative for Educational Reform*. Washington, DC: US Department of Education.
Niaz, M. 2011. *Innovating Science Teacher Education: A History and Philosophy of Science Perspective*. London: Routledge.
Nola, R., and Sankey, H. 2007. *Theories of Scientific Method*. Montreal: McGill-Queen's University Press.
Norton, D. F. (ed.). 1993. *The Cambridge Companion to Hume*. Cambridge, UK: Cambridge University Press.
NRC (National Research Council). 1996. *National Science Education Standards*. Washington, DC: National Academies Press.
1997. *Science Teaching Reconsidered: A Handbook*. Washington, DC: National Academies Press.
1999. *Transforming Undergraduate Education in Science, Mathematics, Engineering, and Technology*. Washington, DC: National Academies Press.
2001. *Educating Teachers of Science, Mathematics, and Technology: New Practices for the New Millennium*. Washington, DC: National Academies Press.
2012. *A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academies Press.

NSF (National Science Foundation). 1996. *Shaping the Future: New Expectations for Undergraduate Education in Science, Mathematics, Engineering, and Technology*. Arlington, VA: National Science Foundation.

NSTA (National Science Teachers Association). 1995. *A High School Framework for National Science Education Standards*. Arlington, VA: National Science Teachers Association.

O'Hear, A. 1989. *An Introduction to the Philosophy of Science*. Oxford, UK: Oxford University Press.

Oreskes, N., Stainforth, D. A., and Smith, L. A. 2010. Adaptation to global warming: Do climate models tell us what we need to know? *Philosophy of Science* 77: 1012–1028.

Palmer, H. 1985. *Presupposition and Transcendental Inference*. New York: St. Martin's Press.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*, 2nd edn. Cambridge, UK: Cambridge University Press.

Piepho, H.-P., and Gauch, H. G. 2001. Marker pair selection for mapping quantitative trait loci. *Genetics* 157:433–444.

Pirie, M. 2006. *How to Win Every Argument: The Use and Abuse of Logic*. New York: Continuum.

Polanyi, M. 1962. *Personal Knowledge: Towards a Post-Critical Philosophy*. Chicago, IL: University of Chicago Press.

Popper, K. R. 1945. *The Open Society and Its Enemies*. London: Routledge & Kegan Paul.

1968. *The Logic of Scientific Discovery*. New York: Harper & Row. (Originally published 1959.)

1974. *Conjectures and Refutations: The Growth of Scientific Knowledge*, 5th edn. New York: Harper & Row.

1979. *Objective Knowledge: An Evolutionary Approach*. Oxford, UK: Oxford University Press.

Potter, M. 2000. *Reason's Nearest Kin: Philosophies of Arithmetic from Kant to Carnap*. Oxford, UK: Oxford University Press.

Raftery, A. E. 1995. Bayesian model selection in social research. *Sociological Methodology* 25:111–163.

Reif, F. 1995. Understanding and teaching important scientific thought processes. *Journal of Science Education and Technology* 4:261–282.

Resnik, D. B. 1998. *The Ethics of Science: An Introduction*. London: Routledge.

Robert, C. P. 2007. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, 2nd edn. New York: Springer.

Robinson, J. T. 1998. Science teaching and the nature of science. *Science & Education* 7:617–634. (Reprint of 1965 article.)

Rosenberg, A. 2012. *Philosophy of Science: A Contemporary Introduction*, 3rd edn. London: Routledge.

Rosenthal-Schneider, I. 1980. *Reality and Scientific Truth: Discussions with Einstein, von Laue, and Planck*. Detroit, MI: Wayne State University Press.

Rothman, K. J., Greenland, S., Poole, C., and Lash, T. L. 2008. Causation and causal inference. In: Rothman, K. J., Greenland, S., and Lash, T. L. (eds.). *Modern Epidemiology*, 3rd edn., pp. 5–31. New York: Wolters Kluwer.

Ryder, J., and Leach, J. 1999. University science students' experiences of investigative project work and their images of science. *International Journal of Science Education* 21:945–956.

Sadler, T. D., Burgin, S., McKinney, L., and Ponjuan, L. 2010. Learning science through research apprenticeships: A critical review of the literature. *Journal of Research in Science Teaching* 47:235–256.

Salmon, W. C. 1967. *The Foundations of Scientific Inference*. Pittsburgh, PA: University of Pittsburgh Press.

1984. *Logic*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.

Sandel, M. J. 2009. *Justice: What's the Right Thing to Do?* New York: Farrar, Straus and Giroux.

Savitz, N. V., and Raudenbush, S. W. 2009. Exploiting spatial dependence to improve measurement of neighborhood social processes. *Sociological Methodology* 39:151–183.

Schilpp, P. A. (ed.). 1951. *Albert Einstein: Philosopher-Scientist*. New York: Tudor Publishing.

(ed.). 1974. *The Philosophy of Karl Popper*, 2 vols. Evanston, IL: Library of Living Philosophers.

Schulze, W. 2009. Field experiments in food and resource economics research: Discussion. *American Journal of Agricultural Economics* 91:1279–1280.

Schupbach, J. N. 2005. On a Bayesian analysis of the virtue of unification. *Philosophy of Science* 72:594–607.

Schupbach, J. N., and Sprenger, J. 2011. The logic of explanatory power. *Philosophy of Science* 78:105–127.

Schwartz, R. S., Lederman, N. G., and Crawford, B. A. 2004. Developing views of nature of science in an authentic context: An explicit approach to bridging the gap between nature of science and scientific inquiry. *Science Education* 88:610–645.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6:461–464.

Seghouane, A.-K. 2009. Model selection criteria for image restoration. *IEEE Transactions on Neural Networks* 20:1357–1363.

Seker, H., and Welsh, L. C. 2006. The use of history of mechanics in teaching motion and force units. *Science & Education* 15:55–89.

Shermer, M. 2004. God's number is up. *Scientific American* 291(1):46.

Simon, H. A. 1962. The architecture of complexity. *Proceedings of the American Philosophical Society* 106:467–482.

Snow, C. P. 1993. *The Two Cultures and the Scientific Revolution*. Cambridge, UK: Cambridge University Press. (Originally published 1959.)

Sokal, A. 1996. Transgressing the boundaries: Toward a transformative hermeneutics of quantum gravity. *Social Text* 46/47:217–252.

2008. *Beyond the Hoax: Science, Philosophy, and Culture*. Oxford, UK: Oxford University Press.

Souders, T. M., and Stenbakken, G. N. 1991. Cutting the high cost of testing. *IEEE Spectrum* 28(3):48–51.

Sprenger, J. 2009. Evidence and experimental design in sequential trials. *Philosophy of Science* 76:637–649.

Stalker, D. (ed.). 1994. *Grue! The New Riddle of Induction*. La Salle, IL: Open Court.
Stehr, N., and Meja, V. (eds.) 2005. *Society & Knowledge: Contemporary Perspectives in the Sociology of Knowledge & Science*, 2nd edn. New Brunswick, NJ: Transaction Publishers.
Stein, C. 1955. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1:197–206.
Stewart, I. 1996. The interrogator's fallacy. *Scientific American* 275(3):172–175.
Stirzaker, D. 1994. *Elementary Probability*. Cambridge, UK: Cambridge University Press.
Stove, D. C. 1982. *Popper and After: Four Modern Irrationalists*. Elmsford Park, NY: Pergamon.
Swinburne, R. 1997. *Simplicity as Evidence of Truth*. Milwaukee, WI: Marquette University Press.
(ed). 2002. *Bayes's Theorem*. Oxford, UK: Oxford University Press.
2003. *The Resurrection of God Incarnate*. Oxford, UK: Oxford University Press.
Taper, M. L., and Lele, S. R. (eds.). 2004. *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*. Chicago, IL: University of Chicago Press.
Theocharis, T., and Psimopoulos, M. 1987. Where science has gone wrong. *Nature* 329:595–598.
Thompson, W. C., and Schumann, E. L. 1987. Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior* 11:167–187.
Tobin, K. (ed.). 1993. *The Practice of Constructivism in Science Education*. Hillsdale, NJ: Lawrence Erlbaum.
Trigg, R. 1980. *Reality at Risk: A Defence of Realism in Philosophy and the Sciences*. Totowa, NJ: Barnes & Noble.
1993. *Rationality and Science: Can Science Explain Everything?* Oxford, UK: Blackwell.
Turner, M. 2010. No miracle in the universe. *Nature* 467:657–658.
Urbach, P. 1987. *Francis Bacon's Philosophy of Science*. La Salle, IL: Open Court.
Urhahne, D., Kremer, K., and Mayer, J. 2011. Conceptions of the nature of science – are they general or context specific? *International Journal of Science and Mathematics Education* 9:707–730.
Wade, N. 1977. Thomas S. Kuhn: Revolutionary theorist of science. *Science* 197:143–145.
Waldrop, M. M. 2011. Faith in science. *Nature* 470:323–325.
Warren, B., Ballenger, C., Ognowski, M., Rosebery, A. S., and Hudicourt-Barnes, J. 2001. Rethinking diversity in learning science: The logic of everyday sense-making. *Journal of Research in Science Teaching* 38:529–552.
Weaver, R. M. 1948. *Ideas Have Consequences*. Chicago, IL: University of Chicago Press.
Weisheipl, J. A. (ed.). 1980. *Albertus Magnus and the Sciences*. Toronto: Pontifical Institute of Mediaeval Studies.
Whitehead, A. N. 1925. *Science and the Modern World*. Cambridge, UK: Cambridge University Press.
Whitehead, A. N., and Russell, B. 1910–13. *Principia Mathematica*, 3 vols. Cambridge, UK: Cambridge University Press.

Williams, L. P., and Steffens, H. J. 1978. *The History of Science in Western Civilization. Vol. II of The Scientific Revolution.* Lanham, MD: University Press of America.

Wolpert, L. 1993. *The Unnatural Nature of Science.* Cambridge, MA: Harvard University Press.

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation.* Oxford, UK: Oxford University Press.

Woolhouse, R. S. 1988. *The Empiricists.* Oxford, UK: Oxford University Press.

Xie, Y. 2002. Comment: The essential tension between parsimony and accuracy. *Sociological Methodology* 28:231–236.

Yang, C.-C. 2007. Confirmatory and structural categorical latent variables models. *Quality & Quantity* 41:831–849.