6

# Science's powers and limits

What are science's powers and limits? That is, where is the boundary between what science is and is not able to discover? The American Association for the Advancement of Science has identified that issue as a critical component of science literacy: "Being liberally educated requires an awareness not only of the power of scientific knowledge but also of its limitations," and learning science's limits "should be a goal in all science courses" (AAAS 1990:20–21). The National Research Council concurs: "Students should develop an understanding of what science is, what science is not, what science can and cannot do, and how science contributes to culture" (NRC 1996:21).

People's motivations for exploring the limits of science can easily be misconstrued, so they should be made clear from the outset. Unfortunately, for some authors writing about science's limits, the motivation has been to exaggerate the limitations in order to cut science down, support antiscientific sentiments, or make more room for philosophy or religion. For others, the motivation has been to downplay science's limitations in order to enthrone science as the one and only source of real knowledge and truth. Neither of those excesses represents my intentions. I do not intend to fabricate specious problems to shrink science's domain, nor do I intend to ignore actual limitations to aggrandize science's claims. Rather, the motivation here is to characterize the actual boundary between what science can do and cannot do. One of the principal determinants of that boundary is the topic of this book, the scientific method.

## Rather obvious limitations

Several limitations of science are rather obvious and hence are not controversial. The most obvious limitation is that scientists will never observe, know, and explain everything about even science's own domain, the physical world. The Heisenberg uncertainty principle, Gödel's theorem, and chaos theory set fundamental limits.

Besides these fundamental limits, there are also practical and financial limits. "Today, the costs of doing scientific work are met by public and corporate funds. Often, major areas of scientific endeavor are determined by the mission-oriented goals of government, industry, and the corporations that provide funds, which differ from the goals of science" (AAAS 1990:21).

The most striking limitation of science, already discussed in Chapter 5, is that science cannot prove its presuppositions. Nor can science appeal to philosophy to do this job on its behalf. Rather, science's presuppositions of a real and comprehensible world – as well as philosophy's presuppositions of the same – are legitimated by an appeal to rudimentary common sense followed by philosophical reflection.

However, the remainder of this chapter explores the powers and limits of science that are not especially obvious. Science's capacity to address big worldview questions is important but controversial. And an integrally related matter is the role of the humanities and the influence of individual experience on worldview convictions. A neglected topic meriting attention is science's power to enhance personal character and experiences of life.

## The sciences and worldviews

Can science reach farther than its ordinary investigations of galaxies, flowers, bacteria, electrons, and such? Can science also tackle life's big questions, such as whether God exists and whether the universe is purposeful? This is the most complex – and perhaps the most significant – aspect of the boundary between science's powers and limits.

Life's grand questions could be termed religious or philosophical or worldview questions. But a single principal term is convenient and the rather broad term *worldview* is chosen here. A worldview sums up a person's basic beliefs about the world and life. The following account draws heavily from Gauch (2009a, 2009b).

Whether worldview implications are part of science's legitimate business is controversial. Nevertheless, the mainstream view, as represented by the AAAS, is that one of science's important ambitions is contributing to a meaningful worldview. "Science is one of the liberal arts" and "the ultimate goal of liberal education" is the "lifelong quest for knowledge of self and nature," including the quest "to seek meaning in life" and to achieve a "unity of knowledge" (AAAS 1990:xi, 12, 21). AAAS position papers offer numerous, mostly helpful perspectives on religion, God, the Bible, clergy, prayer, and miracles. The Dialogue on Science, Ethics, and Religion (DoSER) program of the AAAS offers ongoing events and publications.

The AAAS regards science's influence on worldviews not only as a desirable quest but also a historical reality. "The knowledge it [science] generates

sometimes forces us to change—even discard—beliefs we have long held about ourselves and our significance in the grand scheme of things. The revolutions that we associate with Newton, Darwin, and Lyell have had as much to do with our sense of humanity as they do with our knowledge of the earth and its inhabitants. . . . Becoming aware of the impact of scientific and technological developments on human beliefs and feelings should be part of everyone's science education" (AAAS 1989:134). Likewise, "Scientific ideas not only influence the nature of scientific research, but also influence—and are influenced by—the wider world of ideas as well. For example, the scientific ideas of Copernicus, Newton, and Darwin . . . both altered the direction of scientific inquiry and influenced religious, philosophical, and social thought" (AAAS 1990:24).

But, unfortunately, on the specific worldview question of life's purposes, AAAS position papers are inconsistent. On the one hand, they say that science *does not* answer the big question about purposes: "There are many matters that cannot usefully be examined in a scientific way. There are, for instance, beliefs that—by their very nature—cannot be proved or disproved (such as the existence of supernatural powers and beings, or the true purposes of life)" (AAAS 1989:26). On the other hand, it is most perplexing that another AAAS position paper claims that science *does* answer this question: "There can be no understanding of science without understanding change and the fact that we live in a directional, though not teleological, universe" (AAAS 1990:xiii; also see p. 24). Now "teleological" just means "purposeful," so here the AAAS is boldly declaring, without any argumentation or evidence, that we live in a purposeless universe. Consequently, this is one of those rare instances in which AAAS statements have not provided reliable guidance because they are contradictory.

Science's powers and limits as regards ambitious worldview inquiries depend not only on science's method but also on social conventions that define science's boundaries and interests. A social convention prevalent in contemporary science, *methodological naturalism*, limits science's interests and explanations to natural things and events, not supernatural entities such as God or angels. Methodological naturalism has roots in antiquity with Thales (*c.* 624–546 BC) and others. Subsequently, medieval scholars emphasized pushing their understanding of natural causes to its limits (Lindberg 2007:240–241; Ronald L. Numbers, in Lindberg and Numbers 2003:265–285). But the name "methodological naturalism" is of recent origin, only three decades ago.

Methodological naturalism contrasts with metaphysical or *ontological naturalism* that asserts natural entities exist but nothing is supernatural, as claimed by atheists. Hence, methodological naturalism does not deny that the supernatural exists but rather stipulates that it is outside science's purview. Unfortunately, methodological naturalism is sometimes confused with ontological naturalism. To insist that science obeys methodological naturalism *and* that science supports atheism is to get high marks for enthusiasm but low marks for logic.

Many worldview matters might seem to reside within science's limits, rather than its powers, given that methodological naturalism excludes the supernatural. Indeed, questions such as whether God exists and whether the universe is purposeful, which inherently involve the supernatural, are precisely the kinds of questions that are foremost in worldview inquiries.

However, to be realistic, contemporary science is replete with vigorous discussions of worldview matters. For starters, consider the exceptional science books that manage to become bestsellers. The great majority of them are extremely popular *precisely because* they have tremendous worldview import, such as Collins (2006) and Dawkins (2006). Less popular but more academic books also concern science and worldviews, such as Ecklund (2010).

Furthermore, interest in science's worldview import is a minor but consistent element in mainstream science journals. For instance, religious experience provides one of the standard arguments for theism, but in *American Scientist*, psychologist Jesse Bering (2006) attempted to explain away belief in a deity or an afterlife as a spurious evolutionary by-product of our useful abilities to reason about the minds of others. Likewise, Michael Shermer, the editor of *Skeptic*, has a monthly column in *Scientific American* with provocative items such as "God's number is up" (Shermer 2004). Also, survey results on the religious convictions of scientists were published in *Science* (Easterbrook 1997), and significant commentary on science and religion was provided in *Nature* (Turner 2010; Grayling 2011; Waldrop 2011). To gauge the extent of worldview interests in mainstream science, an interesting little exercise is to visit the websites of journals such as *Nature* and *Science* and search for "religion" to see how many thousands of hits result.

Hence, contemporary scientific practice is far from a consistent and convincing implementation of methodological naturalism. Nor is the present scene uncharacteristic, given the broad interests of natural philosophers (now known as "scientists" since around 1850) in ancient, medieval, and modern times. Of course, methodological naturalism is characteristic of routine scientific investigations, such as sequencing the genome of the virus that causes the common cold, but that does not necessarily mean that it extends to every last scientific interest or publication.

Whereas mainstream science can and does have some worldview import, prominent variants of fringe science are problematic, particularly scientism and skepticism. They are opposite errors. At the one extreme, scientism says that only hard, no-nonsense science produces all of our dependable, solid truth. At the opposite extreme, skepticism says that science produces no final, settled truth.

Yet, curiously, these opposite errors support exactly the same verdict on any worldview inquiry appealing to empirical and public evidence. On the one hand, scientism automatically and breezily dismisses any worldview arguments coming from philosophy, theology, or any other discipline in the humanities

because such disciplines lack the validity and authority that science alone possesses. On the other hand, after skepticism has already judged all science to be awash in uncertainty and tentativeness, ambitious worldview inquiries are bound to receive this same verdict of impotence.

Returning to mainstream science, some scientists explore science's worldview import, other scientists exclude worldview issues in the name of methodological naturalism, and still other scientists have no interests or opinions on such matters whatsoever. This diversity of interests and temperaments hardly seems surprising.

## Empirical method in the humanities

This whole book is about scientific method, but this one section is about a broader topic that may be termed *empirical method*, which subsumes scientific method as a special case. Empirical method concerns what can be known by means of empirical and public evidence, regardless of whether that evidence comes from the sciences or the humanities. Any persons interested in pushing empirical and public evidence to its limits must understand the structure and workings of empirical method, not merely scientific method.

The humanities are academic disciplines that study the human condition. They include the classics, languages, literature, history, law, philosophy, religion or theology, and the visual and performing arts. The humanities use a great variety of methods, including some use of empirical and public evidence.

The essence of scientific method is to appeal to empirical and public evidence to gain knowledge of great theoretical and practical value about the physical world. In greater detail than that single sentence can capture, this book's account of scientific method features the PEL model of full disclosure and the justification of truth claims based on that model, as summarized in Figures 5.1 and 5.3 – although this whole book is needed for a reasonably complete account of scientific method. But, clearly, empirical and public evidence also has roles in the humanities. Especially when empirical evidence is used in ambitious worldview inquiries, as contrasted with routine scientific or technological investigations, the combined perspectives of the sciences and the humanities yield the most reliable and beneficial results.

This section's extremely brief account of empirical method is relevant in this book on scientific method for at least three reasons. First, understanding how public evidence and standard reasoning support truth claims in multiple contexts across the sciences and the humanities gives students their best chance of deeply understanding rationality within science itself. Comparing and contrasting stimulates real comprehension. Second, the AAAS (1990) vision of science as a liberal art calls for a humanities-rich understanding of science, which is promoted greatly by grasping the empirical method that spans the sciences

and the humanities. Third, the scientism that is decisively renounced by mainstream science, but still finds frequent expression especially at the popular level, is best discredited by conscious awareness of projects in the humanities that also appeal to empirical evidence.

All of the disciplines in the humanities contribute to a meaningful worldview. But among these many academic disciplines, natural theology is a prominent example of using empirical method to address worldview questions by means of public evidence.

The article on natural theology by MacDonald (1998) in the *Routledge Encyclopedia of Philosophy* characterizes this discipline. "Natural theology aims at establishing truths or acquiring knowledge about God (or divine matters generally) using only our natural cognitive resources." He further explained that "The phrase 'our natural cognitive resources' identifies both the methods and data for natural theology: it relies on standard techniques of reasoning and facts or truths in principle available to all human beings just in virtue of their possessing reason and sense perception." Natural theology considers arguments both for and against theism, with proponents of diverse perspectives sharing a common impartial methodology.

The implicit contrast is with revealed theology, which instead relies on a revelation or scripture taken as authoritative or inspired within a given faith community. However, a scripture may have some contents and aspects that are verifiable independently with public evidence, so the relationship between natural and revealed theology is one of partial overlap.

Natural theology may be completely unknown to many students and professionals in the sciences. But this unfamiliarity does not negate the existence of this vigorous academic discipline, nor negate natural theology's character as a discipline that relies on empirical and public evidence. Two resources on natural theology may be mentioned for those who are interested. *The Blackwell Companion to Natural Theology* provides a recent and scholarly overview of natural theology (Craig and Moreland 2009). Its chapters review the ontological, cosmological, and moral arguments and the arguments from evil, consciousness, reason, religious experience, and miracles. The ongoing Gifford lectures on natural theology – endowed by Lord Gifford more than a century ago in Scotland's four ancient universities – are frequently published in readily available books.

Gifford lectures by eminent scientists, theologians, philosophers, and other scholars engage an astonishing and intriguing diversity of arguments and evidence. These renowned lectures on natural theology have included scientists Simon Conway Morris, Richard Dawkins, Freeman Dyson, Sir John Eccles, Sir Arthur Eddington, Werner Heisenberg, Michael Polanyi, Martin Rees, and Carl Sagan; theologians Karl Barth, Rudolf Bultmann, Stanley Hauerwas, Jurgen Moltmann, Reinhold Niebuhr, Albert Schweitzer, and Paul Tillich; scientist-theologians Ian Barbour, Stanley Jaki, and Sir John Polkinghorne; philosophers

Marilyn Adams, Sir Alfred Ayer, John Dewey, Antony Flew, Étienne Gilson, Alasdair MacIntyre, Mary Midgley, Alvin Plantinga, Paul Ricoeur, Eleonore Stump, Richard Swinburne, and Alfred Whitehead; and scholars Noam Chomsky, Frederick Copleston, Jaroslav Pelikan, and Arnold Toynbee.

The literature on natural theology – both historical and contemporary – is largely from prestigious academic publishers, and it is simply enormous. Evaluation of natural theology's empirical evidence is outside the scope of this book and it requires considerable effort. By stark contrast, evaluation of natural theology's empirical method is within the scope of this book on scientific method for the three reasons indicated near the beginning of this section, and this evaluation is easy work. It requires merely one longish paragraph, as follows.

The PEL model, which applies to all disciplines and inquiries using empirical and public evidence to support truth claims, specifies three requirements. (1) *Appropriate Presuppositions.* MacDonald's definition of natural theology does not mention presuppositions explicitly, but the context makes two things abundantly clear. On the one hand, natural theology's arguments support conclusions either for or against theistic beliefs, so avoidance of circular reasoning necessarily prohibits natural theology's presuppositions from containing any worldview distinctives. On the other hand, "facts or truths in principle available to all human beings just in virtue of their possessing reason and sense perception" just is public and empirical evidence. Accordingly, like natural science, natural theology must also presuppose the existence and comprehensibility of the physical world. Hence, natural science and natural theology have identical presuppositions. (2) *Admissible and Relevant Evidence.* The admissibility of empirical evidence depends on a methodological consideration, namely, appropriate presuppositions, as already mentioned. And the relevance of empirical evidence depends on whether a given item or collection of admissible evidence bears differentially on the credibilities of the competing hypotheses. To count as relevant evidence in public discourse, the evidence must constitute facts established to everyone's satisfaction, and the interpretation of the evidence must also be settled, which involves agreement over how likely (at least approximately) the observed facts would be were each of the hypotheses true. That is, disputes concern which worldview hypothesis is true or likely, but not the facts, and not the interpretations of the facts. Relevance must be judged on a case-by-case basis and hence is a matter for detailed empirical investigation, rather than a methodological consideration to be resolved by a single decision yielding a comprehensive verdict. (3) *Standard and Impartial Logic.* The logic that natural theology uses "relies on standard techniques of reasoning." The implicit contrast is with special pleading that biases an argument toward the favored conclusion. Natural theology uses the same sorts of deductive and inductive logic as natural science. Logic is explored in the following three chapters, including Bayesian inference that is used extensively in natural theology.

Hypothesis tests using Bayesian methods treat all hypotheses symmetrically and impartially. As will be explained in following chapters, an exceedingly strong conclusion can emerge when the weight of the evidence grows exponentially with its amount. Some arguments in natural theology exemplify this particularly favorable situation. In review, reasons that count across worldviews satisfy three necessary and sufficient conditions: appropriate presuppositions, admissible and relevant evidence, and standard and impartial logic. Natural theology's methodology assures, once and for all, that natural theology has appropriate presuppositions, admissible evidence, and standard and impartial logic. That leaves only the relevance of the evidence for testing specified hypotheses to be judged on a case-by-case basis by means of careful empirical investigation.

Avoidance of circular reasoning is crucial for applications of empirical method in worldview inquiries. Unfortunately, circular reasoning can take much more subtle forms than its obvious archetype, "*X*; therefore *X*." For a first example of subtle circular reasoning, consider the question: Does evolution show that life emerged from random mutations and processes within a purposeless universe? Atheists or agnostics such as Richard Dawkins (1996, 2006) typically presume that random processes like gene mutations must be purposeless. But theist Francis Collins (2006:205, also see 80–82) believes in a sovereign God who inhabits eternity, so "God could be completely and intimately involved in the creation of all species, while from our perspective, limited as it is by the tyranny of time, this would appear a random and undirected process." Hence, there can be agreement about the facts of random mutations and yet disagreement about the interpretation of those facts as regards purposelessness. Until the interpretation of these facts has been settled in a manner that counts across worldviews, any assertion that randomness implies purposelessness constitutes subtle circular reasoning. Why? That implication of purposelessness depends crucially on a particular and supportive worldview, atheism, that is only one of the worldviews included in a conversation taking place in natural theology.

For a second and final example of subtle circular reasoning, consider the question: Can science explain everything? The claim that everything has a scientific, natural explanation has been a popular argument for atheism at least since medieval times. Thomas Aquinas (*c.* 1225–1274) expressed this objection to theism quite concisely: "it seems that we can fully account for everything we observe in the world while assuming that God does not exist" (Davies and Leftow 2006:24). But exactly what is this "everything" that science explains? Scientists in particular and people in general disagree about this "everything" that has happened in our world, largely because of worldview differences. For instance, an interesting exchange between Richard Dawkins, identified as a biologist and "an agnostic leaning toward atheism," and Simon Conway Morris, an evolutionary paleontologist and a Christian, was reported in *Scientific American* (Horgan 2005). Dawkins thought that neither the fine-tuning of the universe

nor the origin of life requires an explanation involving God, whereas many theists judge otherwise. But Conway Morris "asserted that the resurrection and other miracles attributed to Christ were 'historically verifiable'," whereas atheists typically deny that such miracles really happened. Consequently, if an argument for either theism or atheism presupposes a particular and controversial account of this "everything" that has happened in our world and then claims that success in explaining "everything" supports this same worldview, then such an argument is merely a subtle instance of the argument form, "*X*; therefore *X*."

These two examples might prompt a suspicion that all arguments in natural theology, if inspected carefully enough, would reduce to circular reasoning. However, an inference from merely two examples, intentionally selected to illustrate potential problems, to a universal verdict on natural theology constitutes singularly bad inductive logic. The intent here is to stimulate careful assessment, not to justify breezy dismissal.

Historically, the weaker of science or theology often sought support from the stronger: "With the benefit of hindsight we can now see that over the course of the past 150 years a remarkable reversal has taken place. Whereas once the investigation of nature had derived status from its intimate connections with the more elevated disciplines of ethics and theology, increasingly during the twentieth century these latter disciplines have humbly sought associations with science in order to bask in its reflected glory" (Peter Harrison, in Dixon, Cantor, and Pumfrey 2010:28). Nevertheless, whatever legitimacy and success natural theology may have is *not* derived from its similarities with natural science, nor the reverse. "Reason interpreting experience uses many different methods, depending on the subject-matter and the point of view, but they all throw light on one another. Science, then, is not to be confused with other modes of thought, but neither is it to be entirely divorced from them" (Caldin 1949:135). Indeed, it is by understanding rational procedure in multiple instances, with each legitimated on its own merits, that one can best understand rationality within any of its applications.

Besides natural theology, other humanities also apply public and empirical evidence to worldview inquiries, including some arguments in philosophy. And because some religions or worldviews are based substantially on historical events, historical and archaeological evidence can have worldview import. On the other hand, literature, music, and art contribute greatly to cultures and worldviews, but not particularly by way of empirical evidence bearing on worldview hypotheses. In the special case of a scientific or historical inquiry that is especially rich in worldview import, at least as some scholars see it, philosophical and statistical analysis is often essential for a proper assessment of the bearing of the evidence on competing worldview hypotheses, including avoidance of subtle circular reasoning. The principal requirement for any worldview inquiry appealing to public and empirical evidence, whether it be pursued in natural theology or science or history, is that the action be in public evidence, not controversial presuppositions or biased logic. The very fact that

ordinarily worldviews are highly comprehensive tends to implicate multiple possibilities for relevant evidence, so cumulative cases with multiple arguments are common. The inherent strength of a cumulative case, however, comes at the risk of diffuse and rambling argumentation with little action in any one spot. Consequently, a cumulative case is more engaging if at least one of its lines of argumentation is strong, even when considered singly.

In conclusion, mainstream science favors, and historical review exemplifies, science's contribution to a meaningful worldview. But empirical and public evidence from the humanities and sciences together is far more informative than from the sciences alone. The reward for the scientist who perceives scientific method to be an instance of empirical method more generally is the liberty to put empirical evidence to greater use.

## Individual experience and worldviews

The preceding two sections concerned empirical and public evidence from the sciences and the humanities. But public evidence is not the sum total of influences on worldview convictions. People are also influenced by their individual experience, including experience that would not ordinarily count as empirical and public evidence.

For example, consider personal beliefs about whether miracles occur, which can influence worldview convictions substantially. To be clear, what is meant by miracles here is real, decidedly supernatural miracles – not the "miracle" of seeing one's own child born or the "miracle" of getting that dream job. Many persons believe in miracles, either from direct observation or from dependable reports from trusted family and friends, as well as from historical miracle reports in a scripture that is trusted and authoritative within a given religious tradition. And many other persons have encountered nothing whatsoever that seems beyond the ordinary workings of the physical world.

Because worldview convictions are so controversial within the scientific community (Easterbrook 1997; Larson and Witham 1999; Ecklund 2010), it is inappropriate for scientific organizations to take positions on which worldview is true. Furthermore, only scientific evidence is within the provenance and competence of scientific organizations, and yet many scholars, including many scientists, believe for good reasons that a wider survey than science alone can offer is required to reach the most reliable and robust conclusions about worldviews.

On the other hand, because mainstream science asserts that science contributes to a meaningful worldview, it is appropriate for individual scientists to argue that scientific evidence supports a particular worldview. When the worldview convictions of such scientists have also been influenced by the humanities, individual experience, or other significant factors, readers of their arguments will benefit from getting the whole story.

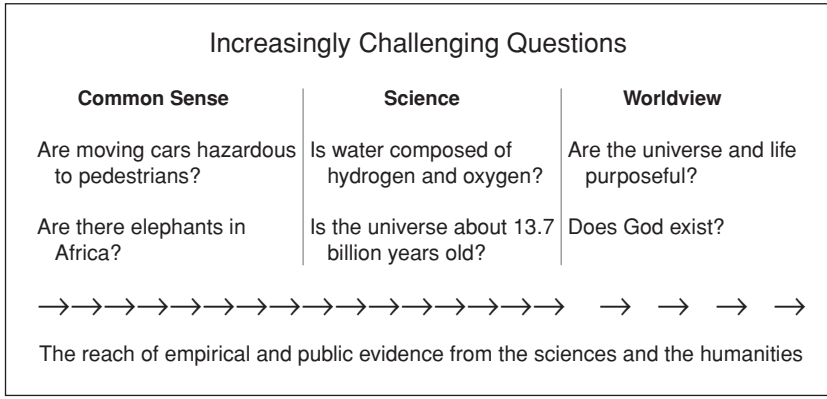


**Figure 6.1 Increasingly challenging questions in the realms of common sense, science, and worldviews. Mainstream science presumes the competence of common sense to answer rudimentary questions, such as those listed in this figure, and affirms the competence of science to answer more difficult questions. But the scientific community lacks consensus on whether empirical and public evidence from the sciences and the humanities can answer challenging worldview questions. Hence, the reach of such evidence is depicted by continuous arrows extending through common sense and science, but by dashed arrows thereafter. All scientists follow along the continuous arrows, whereas only some scientists continue along the dashed arrows.**

To understand the worldview diversity among individuals within the scientific community, a simple but helpful insight is that increasingly challenging questions arise as one progresses from common sense to science to worldview questions, as depicted in Figure 6.1. The underlying issue is the reach of empirical and public evidence from the sciences and the humanities. Such evidence could be of interest for various reasons. Some persons, whether a scientist or not, may think that empirical evidence is the only sort of evidence that really counts. Other persons, especially those with interests in the humanities, may have a broader conception of the sources of knowledge. In either case, a person may want to push empirical and public evidence to its limits, addressing questions as challenging as possible.

Progressing from left to right in this figure, the most rudimentary questions are in the realm of common sense at the left. One who gets that far with empirical and public evidence, having rejected radical skepticism, might well feel encouraged to take the next step: attempting more difficult questions in the realm of science. If that attempt fails, trying even more challenging worldview questions is bound to be futile. But if that attempt succeeds, one might well want to take the next step: attempting yet harder worldview questions, especially by engaging evidence from both the sciences and the humanities.

At its best, the conversation among individual scientists having diverse perspectives on the limits of empirical knowledge is significant, erudite, respectful, and fruitful. What an understanding of scientific method can contribute to that conversation, by drawing on the PEL model, is complete clarity that whatever support science may give to a specific worldview originates from admissible and relevant evidence, rather than from science's worldview-independent presuppositions and impartial logic. What an understanding of science's powers and limits can contribute is a perspective on the liberal art of science that appreciates the combined strength of the sciences and the humanities when tackling ambitious worldview questions, especially questions that methodological naturalism puts outside science's purview. And what a proper understanding of methodological naturalism can contribute is a stipulatory prohibition against invoking supernatural entities within natural science that (1) is not confused with asserting ontological naturalism, and (2) is not extended thoughtlessly to other disciplines outside natural science, such as natural theology, that have their own questions, evidence, and rules.

## Logical roles and diagnoses

The basic components of scientific reasoning – identified by the PEL model as presuppositions, evidence, logic, and conclusions – represent four different logical roles. Different logical roles interact with worldviews in different ways. A statement's logical role is as important as its content.

The difference between "The universe is purposeless" and "The universe is purposeful" is obvious, marking out a vigorous debate. But equally different are "The universe is purposeless" in the logical role of a presupposition and this same "The universe is purposeless" in the role of a conclusion. As a worldview presupposition going beyond science's legitimate presuppositions, its function would be limited to self-congratulatory discourse among kindred spirits. But as a worldview conclusion from a sound argument with worldview-independent presuppositions and impressive evidence, its audience would be the larger world. Recognizing the importance of a statement's logical role, as well as its content, leads to the following several diagnoses.

If a worldview belief has logical roles as both a presupposition and a conclusion within a given discourse, then the diagnosis is circular reasoning in the service of empty dogmatism.

If an argument is unclear regarding whether its worldview belief has the logical role of a presupposition or a conclusion, then the diagnosis is amateurish discourse.

If an argument for a given worldview belief presumes or asserts that science exclusively is the only source of public and empirical evidence, then the diagnosis is the unmitigated scientism that is roundly repudiated as being outside mainstream science (AAAS 1989:26, 30, 133–135, 1990:24–25).

Finally, if a worldview belief emerges in the logical role of a conclusion from an argument also having appropriate presuppositions, admissible and relevant evidence that is public and empirical, and impartial logic, then the diagnosis is a legitimate argument meriting consideration.

## Review of boundaries

Much could be said about the boundaries between science's powers and limits. The following five concise statements express the essence of these boundaries.

(1) The scientific community can build upon and move beyond common sense in providing much reliable and even certain knowledge about the physical world.
(2) Science cannot explain everything about the physical world because of fundamental and practical limits. Also, it cannot prove its own needed presuppositions.
(3) Science is worldview independent as regards its presuppositions and methods, but scientific evidence, or empirical and public evidence more generally, can have worldview import. Methodological considerations reveal this possibility and historical review demonstrates its actuality.
(4) It is appropriate for individual scientists to argue that scientific evidence supports a particular worldview, or else to claim that such arguments are illegitimate. But it is not appropriate for scientific organizations to advocate particular positions because worldview commitments are highly controversial within the scientific community and because the humanities also offer relevant evidence and arguments outside the competence of scientific organizations.
(5) Considerations that inform worldview choice include (1) empirical and public evidence from the sciences; (2) empirical and public evidence from the humanities, especially natural theology; and (3) the individual experience of a given person that is meaningful for that person, although it may not qualify as empirical and public evidence for the wider world. Accordingly, science has significant but limited competence for addressing worldview questions, including whether God exists and whether the universe is purposeful. The sciences without the humanities are lame, and public evidence without individual experience is dehumanizing.

## Personal rewards from science

The intellectual, technological, and economic benefits from science are widely acknowledged by society. Likewise, the importance of science education for good citizenship in a scientific and technological age is widely appreciated. But

another important value of science receives far too little attention: the personal rewards of science, that is, the beneficial effects of science on scientists' personal character and experiences of life.

As this chapter on science's powers and limits draws toward a close, these powers merit attention. Caldin (1949) explored this topic with rare wisdom and charm, so the following remarks draw much from him. Unfortunately, "the place of science in society is too often considered in the narrow setting of economic welfare alone, so that the potential contribution of science to the growth of the mind and will is under-estimated" (Caldin 1949:155). One reward from science is stimulation of rationality and wisdom:

> Now a knowledge of nature is part of wisdom, and we need it to live properly. . . . Science is, therefore, good "in itself," if by that we mean that it can contribute directly to personal virtue and wisdom; it is not just a means to welfare, but part of welfare itself. . . . Scientific life is a version of life lived according to right reason. . . . Consequently, the practice of science requires both personal integrity and respect for one's colleagues; tolerance for others' opinions and determination to improve one's own; and care not to overstate one's case nor to underrate that of others. . . . By studying science and becoming familiar with that form of rational activity, one is helped to understand rational procedure in general; it becomes easier to grasp the principles of all rational life through practice of one form of it, and so to adapt those principles to other studies and to life in general. Scientific work, in short, should be a school of rational life. (Caldin 1949:133–135)

Still another personal reward from science is cultivation of discipline, character, realism, and humility:

> It is not only the intellect that can be developed by scientific life, but the will as well. Science imposes a discipline that can leave a strong mark on the character as can its stimulation of the intellect. All who have been engaged in scientific research know the need for patience and buoyancy and good humour; science, like all intellectual work, demands (to quote von Hügel) "courage, patience, perseverance, candour, simplicity, self-oblivion, continuous generosity towards others, willing correction of even one's most cherished views." Again, like all learning, science demands a twofold attention, to hard facts and to the synthetic interpretation of them; and so it forbids a man to sink into himself and his selfish claims, and shifts the centre of interest from within himself to outside. But for scientists there is a special and peculiar discipline. Matter is perverse and it is difficult to make it behave as one wants; the technique of experimental investigation is a hard and chastening battle. Experimental findings, too, are often unexpected and compel radical revision of theories hitherto respectable. It is in this contact with "brute fact and iron law" that von Hügel found the basis of a modern and scientific asceticism, and in submission to this discipline that he found the detaching, de-subjectifying force that he believed so necessary to the good life. The constant friction and effort, the submission to the brute facts and iron laws of nature, can give rise to that humility and selflessness and detachment which ought to mark out the scientist. (Caldin 1949:135–136)

## Summary

Understanding the boundary between science's powers and limits is a core component of scientific literacy. Some limitations are rather obvious, such as that science cannot explain everything about the world and that it cannot prove its own needed presuppositions.

Many scientists, philosophers, and other scholars have debated whether helping to inform large worldview issues, such as the purpose of life, is among science's powers or beyond its limits. However, the mainstream position represented by the AAAS is that contributing to a meaningful worldview is both a proper ambition of science and a historical reality of science. But the sciences are not alone in this endeavor. Many disciplines in the humanities also contribute to a meaningful worldview, including philosophy, theology, history, literature, and the visual and performing arts. In addition to public evidence from the sciences and the humanities, individual experience can also inform a person's worldview convictions, even though personal experience may not count as public evidence.

Empirical method uses empirical and public evidence from the sciences and the humanities to reach conclusions that can bear on worldviews. In assessing arguments for or against a given worldview, not only the content but also the logical role of statements matters. An argument merits consideration that presents its worldview in the logical role of a conclusion, emerging from appropriate presuppositions, empirical evidence, and impartial logic.

Among science's powers is a considerable ability to be of benefit to scientists' personal character and experiences of life. The essence of this benefit is the selflessness, detachment, and humility that result from deliberate and outward-looking attention given to the physical world.

## Study questions

(1) The AAAS insists that understanding the boundary between science's powers and limits is a core requirement of scientific literacy. Have you received any instruction on these matters? If so, what was the message, did it make sense, and did it align with position papers on science from the AAAS and NRC? If not, what explanations might you suggest for its absence in your curriculum?

(2) Which components of science – presuppositions or logic or evidence – could potentially have worldview import? Explain all three of your verdicts. What diagnosis results if a worldview belief has logical roles of both a presupposition and a conclusion?

(3) The text argues that worldview convictions can be informed by three sources: the sciences, the humanities, and personal experiences. What is

a significant example of each? What does having three potential resources imply for science's role in forming worldview convictions?

(4) What is the distinction between scientific method and empirical method? How do they differ in their powers and limits, particularly in the range of hypotheses and evidence under consideration?

(5) Regardless of whether you are a student or a professional in the sciences or the humanities, what personal rewards by way of wisdom, discipline, and character have you gained from your experiences with science?

7

# Deductive logic

The preceding five chapters are directed mainly at this book's purpose of cultivating a humanities-rich perspective on science. This is the first of five chapters directed mainly at this book's other purpose of increasing scientific productivity.

Logic is the science of correct reasoning and proof, distinguishing good reasoning from bad. Logic addresses the relationship between premises and conclusions, including the bearing of evidence on hypotheses.

In the context of logic, an "argument" is not a dispute but rather is a structured set of statements in which some statements, the premises, are offered to support or prove others, the conclusions. Many deductive systems, including arithmetic and geometry, are developed on a foundation of logic in the modern and unified vision of mathematics.

Of course, given the simple premises that "All men are mortal" and "Socrates is a man," one trusts scientists to reach the valid conclusion that "Socrates is mortal," even without formal study of logic. But given the more difficult problems that continually arise in science, the rate of logical blunders can increase substantially in the absence of elementary training in logic. Fortunately, most blunders involve a small number of common logical fallacies, so even a little training in logic can produce a remarkable improvement in reasoning skills.

The aim of this chapter differs from that of an ordinary text or course on logic. One short chapter cannot teach logic comprehensively. What it can do, however, is convey an insightful general impression of the nature and structure of deductive logic. Recall that the PEL model introduced in Chapter 5 identifies logic as one of the three essential inputs (along with presuppositions and evidence) required to support scientific conclusions. Consequently, the credibility of science depends on having a logic that is coherent and suitable for investigating the physical world.

This chapter distinguishes the two basic kinds of logic: deductive logic, explained in this chapter, and inductive logic, explored in Chapter 9. One branch of deduction, probability theory, is deferred to the next chapter. The

history of logic is reviewed briefly, followed by basic accounts of propositional logic, predicate logic, and arithmetic. Common logical fallacies are analyzed to refine reasoning skills.

## Deduction and induction

The distinction between deduction and induction can be explained in terms of three interrelated differences. Of these three differences, the one listed first is the fundamental difference, with the others being consequences or elaborations. Custom dictates distinct appellative terms for good deductive and inductive arguments. A deductive argument is valid if the truth of its premises guarantees the truth of its conclusions and is invalid otherwise. An inductive argument is strong if its premises support the truth of its conclusions to a considerable degree and is weak otherwise. The following deductive and inductive arguments, based on Salmon (1984:14), illustrate the three differences.

***Valid Deductive Argument***
Premise 1. Every mammal has a heart.
Premise 2. Every horse is a mammal.
Conclusion. Every horse has a heart.

***Strong Inductive Argument***
Premise 1. Every horse that has been observed has had a heart.
Conclusion. Every horse has a heart.

First, the conclusion of a deductive argument is already contained, usually implicitly, in its premises, whereas the conclusion of an inductive argument goes beyond the information present, even implicitly, in its premises. The technical terms for this difference are that deduction is nonampliative but induction is ampliative. For example, the conclusion of the foregoing deductive argument simply states explicitly, or reformulates, the information already given in its premises. All mammals have hearts according to the first premise, and that includes all horses according to the second premise, so the conclusion follows that every horse has a heart. On the other hand, the conclusion of the foregoing inductive argument contains more information than its premise. The premise refers to some group of horses that have been observed up to the present, whereas the conclusion refers to all horses, observed or not, and past or present or future.

Note that this difference, between ampliative and nonampliative arguments, concerns the relationship between an argument's premises and conclusions, specifically whether or not the conclusions contain more information than the premises. This difference does not pertain to the conclusions as such, considered

in isolation from the premises – indeed, the foregoing two arguments have exactly the same conclusion.

Second, given the truth of all of its premises, the conclusion of a valid deductive argument is true with certainty, whereas even given the truth of all of its premises, the conclusion of an inductive argument is true with at most high probability. This greater certainty of deduction is a direct consequence of its being nonampliative: "The [deductive] conclusion must be true if the premises are true, *because* the conclusion says nothing that was not already stated by the premises" (Salmon 1984:15). The only way that the conclusion of a valid deductive argument can be false is for at least one of its premises to be false. On the other hand, the uncertainty of induction is a consequence of its being ampliative: "It is because the [inductive] conclusion says something not given in the premise that the conclusion might be false even though the premise is true. The additional content of the conclusion might be false, rendering the conclusion as a whole false" (Salmon 1984:15). For example, the foregoing inductive conclusion could be false if some other horse, not among those already observed and mentioned in this argument's premise, were being used for veterinary research and had a mechanical pump rather than a horse heart.

Deductive arguments are either valid or invalid on an all-or-nothing basis because validity does not admit of degrees. But inductive arguments admit of degrees of strength. One inductive argument might support its conclusion with a very high probability, whereas another might be rather weak.

The contrast between deduction's certainties and induction's probabilities can easily be overdrawn, however, as if to imply that induction is second-rate logic compared with deduction. Representing certain truth by a probability of 1 and certain falsehood by 0, an inductive conclusion can have any probability from 0 to 1, including values arbitrarily close to 1 representing certainty of truth (or 0 representing certainty of falsehood). Given abundant evidence, induction can deliver practical certainties, although it cannot deliver absolute certainties.

Third and finally, deduction typically reasons from the general to the specific, whereas induction reasons in the opposite direction, from specific cases to general conclusions. That distinction was prominent in Aristotle's view of scientific method (Losee 2001:5–8) and remains prominent in today's dictionary definitions. For instance, the *Oxford English Dictionary* defines "deduction" as "inference by reasoning from generals to particulars," and it defines "induction" as "The process of inferring a general law or principle from the observation of particular instances." Deduction reasons from a given model to expected data, whereas induction reasons from actual data to an inferred model, as depicted in Figure 7.1.

As encountered in typical scientific reasoning, the "generals" and "particulars" of deduction and induction have different natures and locations. The general models or theories exist in a scientist's mind, whereas the particular instances pertain to physical objects or events that have been observed. Often,

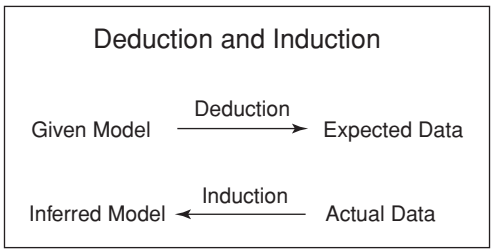


**Figure 7.1 The opposite reasoning directions of deduction and induction. Deduction reasons from the mind to the world, whereas induction reasons from the world to the mind.**

the observations or data comprise a limited sample, but the researchers are interested in the larger population from which the sample was drawn. For instance, a clinical trial may examine a representative sample of persons to reach conclusions pertaining to the whole population of persons suffering from a given disease.

Deduction is neither better than induction nor worse. Rather, they pursue answers to different kinds of questions, with deduction reasoning from a mental model to expected data, and induction reasoning from actual data to a mental model. Both are indispensable for science.

## Historical perspective on deduction

Aristotle (384–322 BC) wrote extensively on logic. Although his works on some topics, including natural science, suffered much neglect until the early 1200s, his corpus on logic, the *Organon* or tool (of reasoning), fared better. Aristotle's logic built on ideas from Socrates and Plato. Epicurean, Stoic, and Pythagorean philosophers also developed logic and mathematics. Besides Greece, there were impressive ancient traditions in logic in Babylon, Egypt, India, and China. Largely because of Augustine's early influence, the Aristotelian tradition came to dominate logic in the West, so that tradition is emphasized here.

In his *Prior Analytics*, Aristotle taught that every belief comes through either deduction or induction. His syllogistic logic is the first deductive system, predating Euclid's geometry. Aristotle proposed an inductive–deductive model of scientific method that features alternation of deductive and inductive steps. This alternation, moving from mental model to physical world and back again, leads scientists to a mind–world correspondence – to truth. In this process, any discrepancy between model and world is to be resolved by adjusting the model to the world because the actual data in the inductive step have priority over the expected data in the deductive step. Assuming that the data are not faulty

or excessively inaccurate, actual data contrary to a model's expectations imply that something is wrong with the model. Adjusting one's model to the world is the basis of scientific realism.

Euclid (*fl. c.* 300 BC) was the great master of geometry. Many truths of geometry were known before Euclid. For example, earlier Babylonians and Egyptians knew that the sum of the interior angles of a triangle equals 180 degrees. But that was known by empirical observation of numerous triangles, followed by inductive generalization. Their version of geometry was a practical art related to surveying, in line with the name "geometry," literally meaning "earth measure."

Euclid's *Elements of Geometry*, in one of the greatest paradigm shifts ever, instead demonstrated those geometrical truths by deduction from several axioms and rules. Euclid's geometry had five postulates concerning geometry, such as that a straight line can be extended in either direction, plus five axioms or "common notions" concerning correct thinking and mathematics in general, such as that the whole is greater than its parts. Euclid's combination of geometrical postulates and logical axioms represented a nascent recognition that logic underlies geometry. Countless theorems can be deduced from Euclid's postulates and axioms, including that the sum of the interior angles of a triangle equals 180 degrees.

Subsequently, non-Euclidean geometries were discovered by Thomas Reid (1710–1796), Nikolai Lobachevsky (1792–1856), János Bolyai (1802–1860), Bernhard Riemann (1826–1866), and others. This rendered Euclid's work *a* geometry rather than *the* geometry. In Reid's alternative geometry, the sum of the interior angles of a triangle equals more than 180 degrees.

Anicius Manlius Severinus Boethius (AD 480–524) translated, from Greek into Latin, many parts of Aristotle's logical works, Porphyry's *Introduction to Aristotle's Logic*, and parts of Euclid's *Elements*. His *On Arithmetic*, based on earlier work by Nicomachus of Gerasa, became the standard text on arithmetic for almost a millennium.

Peter Abelard (*c.* 1079–1142) wrote four books on logic. He and his students, John of Salisbury and Peter Lombard, greatly influenced medieval logic. The use of Arabic numerals was spread into Europe by Alexandre de Villedieu (*fl. c.* 1225), a French Franciscan, John of Halifax (or Sacrobosco, *c.* 1200–1256), an English schoolman, and Leonardo of Pisa (or Fibonacci, *c.* 1180–1250), an Italian mathematician. The modern mind can hardly imagine the tedium of multiplication or division using Roman numerals, or how few persons in medieval Europe could perform what we now regard as elementary calculations.

Albertus Magnus (*c.* 1200–1280) wrote 8,000 pages of commentary on Aristotle, including much logic. He also wrote a commentary on Euclid's *Elements*.

Robert Grosseteste (*c.* 1168–1253) founded the mathematical-scientific tradition at Oxford. He affirmed and refined Aristotle's inductive–deductive

scientific method, which he termed the "Method of Resolution and Composition" for its inductive and deductive components, respectively. Also, his Method of Verification involved deriving the deductive consequences of a theory beyond the original facts on which the theory was based and then observing the actual outcome in a controlled experiment to check the theory's predictions. That method recognized the priority of data over theories, in accord with Aristotle. Grosseteste's Method of Falsification eliminated bad theories or explanations by showing that they imply things known to be false. To increase the chances of eliminating false theories, he recommended that conclusions reached by induction be submitted to the test of further observation or experimentation.

Putting all those methods together, the objective of Grosseteste's new science was to make theory bear on the world and the world bear on theory, thereby bringing theory into correspondence with the world. Grosseteste's scientific method sought to falsify and reject false theories, to confirm and accept true theories, and to discern which kinds of observational or experimental data would help the most in theory evaluation.

There is substantial similarity between Grosseteste's medieval science and modern science. "Modern science owes most of its success to the use of these inductive and experimental procedures, constituting what is often called 'the experimental method'. The . . . modern, systematic understanding of at least the qualitative aspects of this method was created by the philosophers of the West in the thirteenth century. It was they who transformed the Greek geometrical method into the experimental science of the modern world" (Crombie 1962:1). I concur with this assessment that a basically correct and complete scientific method emerged in the thirteenth century.

William of Ockham (*c.* 1285–1347) wrote a substantial logic text, the *Summa logicae*. The principle of parsimony is often called Ockham's razor because of his influential emphasis on this principle. Jean Buridan (*c.* 1295–1358) wrote the *Summulae de dialectica*, a then-modern revision and amplification of the earlier logic text by Peter of Spain (*fl.* first half of the thirteenth century), and two advanced texts, the *Consequentiae* and *Sophismata*.

René Descartes (1596–1650) was the founder of analytic geometry. Blaise Pascal (1623–1662) contributed to projective geometry, arithmetic, combinatorial analysis, probability, and the theory of indivisibles (a forerunner of integral calculus). He developed the first commercial calculating machine. Isaac Newton (1642–1727) and Gottfried Leibniz (1646–1716) invented calculus. Giuseppe Peano (1858–1932) devised axioms for arithmetic.

For millennia, the various branches of deduction – such as logic, arithmetic, and geometry – had been developed as separate and unrelated systems. Early great works aiming to unify logic and mathematics were the brilliant *Grundgesetze der Arithmetik* (*The Basic Laws of Arithmetic*) of Frege (1893) and the monumental *Principia Mathematica* of Whitehead and Russell (1910–13).

**Table 7.1.** Truth-table definitions for negation, conjunction, disjunction, implication, and equality

| Assignments | | Not | And | Or | Implies | Equals |
|---|---|---|---|---|---|---|
| $A$ | $B$ | $\sim B$ | $A \wedge B$ | $A \vee B$ | $A \rightarrow B$ | $A \equiv B$ |
| T | T | F | T | T | T | T |
| T | F | T | F | T | F | F |
| F | T | | F | T | T | F |
| F | F | | F | F | T | T |

## Propositional logic

Propositional logic, also called statement calculus and truth-functional logic, is a rather elementary branch of deductive logic. Nevertheless, it is quite important because it pervades common-sense reasoning and scientific reasoning.

A simple proposition has a subject and a predicate, such as "This apple is red" or "Mary is coming." Propositional logic considers only declarative statements. Accordingly, every simple proposition has the property of having one or the other of two possible truth-values: true (T) and false (F). Note that the truth-value applies to the proposition as a whole, such as "This apple is red" being true for a red apple but false for a green apple. In propositional logic, as introduced in this section, there is no further analysis of the subject and predicate within a proposition. But, in predicate logic, to be explained in the next section, further analysis is undertaken. Hence, predicate logic is more complicated, subsuming propositional logic and adding new concepts and analysis.

Proposition constants represent specific simple propositions and are denoted here by uppercase letters like $A$, $B$, and $C$ (except that T and F are reserved to represent the truth-values true and false). For example, "The barometer falls" can be symbolized by $B$, "It will rain" by $R$, and "It will snow" by $S$. Then, the compound sentence "If the barometer falls, then either it will rain or it will snow" can be expressed by "If $B$, then $R$ or $S$."

The most common connectives or operators are "not," "and," "or," "implies," and "equals." They are also termed negation, conjunction, disjunction, implication, and equality. These five connectives are denoted here by these symbols: "$\sim$," "$\wedge$," "$\vee$," "$\rightarrow$," and "$\equiv$." The meanings of these connectives are specified by a truth table (Table 7.1).

"Not" is a unary operator applied to a single proposition. If $B$ is true, then $\sim B$ is false; and if $B$ is false, then $\sim B$ is true. That is, $B$ and $\sim B$ have opposite truth-values. The other connectives are binary operators applied to two propositions. For example, "$A$ and $B$," also written as "$A \wedge B$," is true when

both $A$ is true and $B$ is true and is false otherwise. Simple propositions can be combined with connectives, such as $B \rightarrow (R \vee S)$ to symbolize the preceding compound proposition about a barometer. Parentheses are added as needed to avoid ambiguity. To simplify expressions, the conventions are adopted that negation has priority over other connectives and applies to the shortest possible sub-expression, and parentheses may be omitted whenever the order makes no difference.

Incidentally, two other logical operators, not already specified in Table 7.1, are important in computer design because they can be implemented with simple transistor circuits. Joint denial of $A$ and $B$, expressed by "Neither $A$ nor $B$" and symbolized by "$A \downarrow B$," equals the negation of "$A$ or $B$" and hence is also named "Nor." An alternative denial of $A$ and $B$, expressed by "Either not $A$ or not $B$" and symbolized by "$A \mid B$," equals the negation of "$A$ and $B$" and hence is also named "Nand." (To avoid potential confusion, note that this symbol "|" instead means "or" in several computer-programming languages.) Remarkably, all of the logical operators in Table 7.1 can be defined or replaced by joint denial alone, or by alternative denial alone. For instance, $\sim A$ is logically equivalent to $A \downarrow A$ or to $A \mid A$. Likewise, $A \wedge B$ is logically equivalent to $(A \downarrow A) \downarrow (B \downarrow B)$ or to $(A \mid B) \mid (A \mid B)$. Consequently, circuits using Nor and Nand operations are extremely useful in computers. Annually, the world produces more transistors than it produces grains of wheat or grains of rice.

Proposition variables stand for simple propositions and are denoted here by lowercase letters like $p$ and $q$. Hence, the variable $p$ could stand for the constant $A$ or $B$ or $C$. Proposition expressions are denoted here by script letters and are formed by one or more applications of two rules: (1) any proposition constant or variable is a proposition expression; and (2) if $\mathcal{A}$ and $\mathcal{B}$ are proposition expressions, then their negations are proposition expressions as well as their being combined by conjunction, disjunction, implication, and equality.

An argument is a structured, finite sequence of proposition expressions, with the last being the conclusion (ordinarily prefaced by the word "therefore" or the symbol "$\therefore$"), and the others the premises. The premises are intended to support or prove the conclusion. For example, *modus ponens* is a valid argument with two premises and one conclusion: $A$; $A$ implies $B$; therefore $B$. Likewise, *modus tollens* is the valid argument: not $B$; $A$ implies $B$; therefore not $A$. Incidentally, the full Latin names are *modus ponendo ponens* meaning "the way that affirms by affirming," and *modus tollendo tollens* meaning "the way that denies by denying." An argument is valid if under every assignment of truth-values to the proposition variables that makes all premises true, the conclusion is also true. Otherwise, the argument is invalid.

There are several methods for proving that an argument is valid or else invalid, as the case may be. Different methods all give the same verdict, but one

### Formal Propositional Logic

(1) Language. The symbols used are as follows: $\sim$, $\rightarrow$, (, ), $p_1$, $p_2$, $p_3$, and so on.
(2) Expressions. A well-formed formula (wff) is formed by one or more applications of two rules. (a) Each $p_i$ is a wff. (b) If $A$ and $B$ are wffs, then $(\sim A)$ and $(A \rightarrow B)$ are wffs.
(3) Axioms. For any wffs $A$, $B$, and $C$, axioms are formed by the following three axiom schemes:
  Axiom Scheme 1. $(A \rightarrow (B \rightarrow A))$
  Axiom Scheme 2. $((A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C)))$
  Axiom Scheme 3. $(((\sim A) \rightarrow (\sim B)) \rightarrow (B \rightarrow A))$
(4) Rule. The rule, *modus ponens*, says from $A$ and $(A \rightarrow B)$, infer $B$.
(5) Interpretation. The symbols "$\sim$" and "$\rightarrow$" are the logical connectives negation and implication, which may have associated parentheses needed to specify the order of operations, and the symbols $p_1$, $p_2$, $p_3$, and so on, represent proposition variables having truth-values of either true or false. A proof is a sequence of wffs $A_1, \ldots, A_N, A$ such that each wff either is an axiom or follows from two previous members of the sequence by application of *modus ponens*. The final wff, $A$, is a theorem.

**Figure 7.2 The elements of formal propositional logic. This logic is specified by its language, expressions, axioms, rule, and interpretation.**

method may be easier to understand or use in a given instance than is another. The conceptually simplest method, directly reflecting the definitions of validity and invalidity, is to construct a truth table to determine whether or not each assignment of truth-values to the argument's proposition variables that makes all premises true also makes the conclusion true. Another method for proving validity is to deduce the argument as a theorem from the axioms and rules. But proof strategies are best left to any standard logic text.

Figure 7.2 presents a formal system for propositional logic, drawing on Hamilton (1978:28). Some liberty has been taken to simplify this presentation. The ordinary letters "$A$," "$B$," and "$C$" in this figure should actually be script letters to represent proposition expressions, not merely proposition constants as denoted by these ordinary letters elsewhere in this section.

Propositional logic is both sound and complete. Basically, this means that its rules are correct and that no additional rules are needed. Propositional logic is also decidable, meaning that any argument can be proven to be valid or else invalid. For example, consider the argument: $A \rightarrow B$; $B \rightarrow (C \vee D)$; $A \wedge \sim C$; $\therefore D$. Is it valid or invalid? Some time and effort are required to render the verdict, which turns out to be that this argument is valid. But before even starting to assess validity, it is already known and guaranteed in advance that the outcome is predetermined by the axioms of propositional logic and the answer is decidable.

## Predicate logic

Predicate logic, also called first-order logic and predicate calculus, subsumes and surpasses prepositional logic. It adds two extensions. First, it distinguishes between a proposition's subject and predicate. In a conventional symbolism, the predicate is denoted by an uppercase letter and the subject is denoted by the following lowercase letter placed within parentheses. For instance, "This apple is red," with its predicate "is red" and subject "This apple," can be symbolized by $R(a)$. Second, predicate logic has the existential quantifier "some" denoted by "$\exists$" and the universal quantifier "all" denoted by "$\forall$." For example, $(\forall x)(A(x))$ means "All $x$ are $A$" and $(\exists x)(A(x))$ means "Some $x$ is $A$."

One kind or subset of predicate logic is syllogistic logic. A familiar example is the argument: All men are mortal; Socrates is a man; therefore, Socrates is mortal. Because of its deeper analysis distinguishing subjects and predicates and its inclusion of existential quantifiers such as "all," predicate logic can analyze this argument and declare this syllogism valid. But the simpler propositional logic cannot express or handle syllogisms.

A formal deductive system for predicate logic is about twice as complicated as the one shown in Figure 7.2 for propositional logic (Hamilton 1978:49–56, 71–72). By contrast, the range of theorems that predicate logic can prove is incomparably greater than the range for propositional logic. Accordingly, predicate logic supplies the powerful logic that lays the foundation upon which other branches of mathematics can be constructed, including arithmetic and probability.

## Arithmetic

In the contemporary vision of deductive systems, numerous branches of mathematics, such as arithmetic, are all built on a foundation of predicate logic. To build a branch of mathematics on logic, two items must be added: an interpretation and some axioms. A formal language is abstract, and an interpretation attaches a particular meaning to some symbols of a formal language, such as arithmetic being about numbers. Additional axioms are needed because often a mathematical statement is true (or false) because of the mathematical meanings of its terms, rather than merely the logical arrangement of its terms. There are both logical truths and mathematical truths. Both require axioms.

Although syllogistic logic was axiomatized by Aristotle and geometry by Euclid more than two millennia ago, arithmetic was axiomatized only just over a century ago in 1889 by Giuseppe Peano. His axioms can be edited in various ways to make them somewhat more transparent or convenient. Figure 7.3 presents a formulation with nine axioms.

Peano Axioms for Arithmetic

1. 0 is a natural number.
2. For every natural number $x$, $x = x$.
3. For all natural numbers $x$ and $y$, if $x = y$, then $y = x$.
4. For all natural numbers $x$, $y$, and $z$, if $x = y$ and $y = z$, then $x = z$.
5. For all $a$ and $b$, if $a$ is a natural number and $a = b$, then $b$ is also a natural number.
6. For every natural number $n$, its successor S($n$) is a natural number.
7. For every natural number $n$, S($n$) = 0 is false.
8. For all natural numbers $m$ and $n$, if S($m$) = S($n$), then $m = n$.
9. If K is a set such that (a) 0 is in K and (b) for every natural number $n$, if $n$ is in K, then S($n$) is in K; then K contains every natural number.

**Figure 7.3 The nine Peano axioms for arithmetic. The first axiom assumes that zero is a natural number, the next four describe the equality relation, the next three describe the successor function (where 1 is the successor of 0, 2 is the successor of 1, and so on), and the last axiom concerns the set of all natural numbers.**

Arithmetic can be developed from the Peano axioms and the inherited predicate logic. For the most part, the meanings of the Peano axioms should be fairly obvious. For instance, axiom 2 says that every number equals itself, and axiom 8 says that if $m + 1 = n + 1$, then $m = n$.

To reiterate an important point about predicate logic from the preceding section in the present context of arithmetic, axioms fix in advance the outcome for subsequent theorems or calculations. For example, is $871 \times 592 = 515432$ correct? A little effort is required to check this calculation, but before even starting, the verdict has been predetermined by the arithmetic axioms. Actually, this calculation is incorrect, the proper value being 515632. Precisely because the rules of arithmetic are fixed before the game begins, arithmetic is meaningful and rational. If different persons could get different sums for $27 + 62$, then in such a world there would be no science, and no banks either.

Many persons may miss the wonder, but Albert Einstein asked "How is it possible that mathematics, a product of human thought that is independent of experience, fits so excellently the objects of physical reality?" (Frank 1957:85). Likewise, Potter (2000:17–18) expressed this wonder specifically as regards arithmetic, remarking that "it is not immediately clear why the properties of abstract objects [numbers] should be relevant to counting physical or mental ones. . . . One has only to reflect on it to realize that this link between experience, language, thought, and the world, which is at the very centre of what it is to be human, is truly remarkable."

Indeed, there is something wonderful about arithmetic's effectiveness. It may be noted, however, that the Peano axioms generate standard arithmetic,

whereas there also exist equally internally coherent but different *non*standard arithmetics. For instance, in standard (Peano) arithmetic $2 + 2 = 4$. But, in the nonstandard ring arithmetic based on the circular and repeating arrangement of integers 0, 1, 2, 3, 0, 1, 2, 3, 0, and so on, the sum of interest becomes $2 + 2 = 0$. Likewise, in the ring arithmetic with just 0, 1, and 2 repeating, the sum becomes $2 + 2 = 1$. All three of these arithmetics are equally internally coherent, although they are also different from each other.

There are occasional practical uses for nonstandard arithmetics or geometries. For example, standard arithmetic says that $11 + 3 = 14$. But an ordinary clock is based on a circular arrangement of its integers from 1 to 12, so this ring arithmetic says that 3 hours after 11 o'clock, the time is 2 o'clock, or $11 + 3 = 2$. (Or, some clocks have instead the integers 1 to 24 written in a circle). For another example, ordinary surveying or earth-measure uses ordinary geometry. But airplane pilots traveling great distances use the non-Euclidean geometry that Reid invented (for studying optics in a roughly spherical eye) to follow the shortest great circle bearing on our spherical earth, thereby saving time and fuel.

But apart from these understandable exceptions, standard logic and arithmetic and geometry prevail in daily life. While properly appreciating the wonder of arithmetic, part of the reason that (standard) arithmetic fits with our experiences in the physical world is that the choice of standard over nonstandard arithmetic has been guided preemptively by our interests and needs as incarnate beings living in the physical world. That is, in choosing an arithmetic (or geometry or whatever), coherence is not our only criterion but also fit with our experiences of the world. Hence, in the mathematical world of coherent arithmetics, one can obtain $2 + 2 = 0$ or $2 + 2 = 1$ or $2 + 2 = 4$. However, in the physical world of actual objects and events, standard arithmetic is uniquely appropriate. Two apples plus two apples equals four apples.

## Common fallacies

Ever since Aristotle's *Sophistical Refutations*, logicians have been providing helpful analyses and classifications of logical fallacies. Furthermore, science educators report that "all the standard logical fallacies, known since Aristotle's day, are routinely committed by science students" (Matthews 2000:331).

There are many fine books and resources on fallacies. But the book by Madsen Pirie, with its generous list of 79 fallacies, is outstanding because of its fun rhetoric in the guise of a naughty sophist. He explained: "This book is intended as a practical guide for those who wish to win arguments. It also teaches how to perpetrate fallacies with mischief at heart and malice aforethought. . . . I have given the reader recommendations on how and where the fallacy may be used to deceive with maximum effect. . . . In the hands of the wrong person this is more of a weapon than a book, and it was written with that wrong person in


**Figure 7.4 The logical fallacy *argumentum ad lapidem*, argument to a stone. Samuel Johnson vigorously kicks a stone, attempting to refute the idea that the physical world does not exist, while an unimpressed George Berkeley observes. (This drawing by Carl R. Whittaker is reproduced with his kind permission.)**

mind" (Pirie 2006:ix–x). This is the book that everyone needs as we set about the all-important business of getting our own way!

The study of fallacies best begins with its opposite, the study of right thinking. Knowing the genuine article makes its counterfeits more obvious. Recalling the PEL model in Chapter 5, the essence of scientific thinking is evidence that is *admissible* relative to the presuppositions and *relevant* relative to the hypotheses, as well as deductive and inductive *logic* to draw conclusions and weigh evidence. The three italicized words emphasize the principal opportunities for defects: inadmissible evidence, irrelevant evidence, and fallacious logic. The fourth and final category of fallacies reviewed in this section involves a personal rather than a procedural defect, failure of will to pursue the truth.

**Inadmissible Evidence.** The *argumentum ad lapidem* (argument to a stone) appeals to inadmissible evidence. This fallacy is named for a famous incident depicted in Figure 7.4. George Berkeley had argued that only minds and ideas

exist, not physical objects and events, as mentioned in Chapter 5. When Dr. Samuel Johnson was told by James Boswell that this argument is impossible to refute, he vigorously kicked a stone, exclaiming "I refute it thus."

But as Pirie (2006:101–104) observed, Johnson was not so much refuting Berkeley's argument as ignoring it. Johnson was presuming a realist interpretation or ontology regarding the empirical evidence provided by sight or feel or sound or any other sense, which is precisely the matter in dispute given Berkeley's idealist ontology. As emphasized in Chapter 5, the existence and comprehensibility of the physical world are presuppositions of mainstream science, not conclusions of science (or philosophy either). To think otherwise is to commit the *argumentum ad lapidem* fallacy. Berkeley could accept that Johnson had an experience of kicking a stone and could even share that experience with him. But Berkeley would not infer from this experience the metaphysical theory that the stone has an independent physical existence. Presuppositions cut deeper than evidence.

**Irrelevant Evidence.** Several fallacies appeal to evidence that is admissible, given the common-sense presuppositions of mainstream science, but that evidence is irrelevant because it fails to bear on the credibilities of the various hypotheses under consideration. One such fallacy is the *argumentum ad hominem* (argument to the man), which attacks the person promoting the disliked idea rather than the idea itself. For instance, a theory could be attacked by saying its proponent is a teacher at a small community college.

Another fallacy is the *red herring*. This draws attention away from the original argument to some other matter that is irrelevant but provides an easier target for refutation.

An alluring fallacy for scientists is *unobtainable perfection*, or at least excessive perfection. This fallacy discredits a result by requiring greater accuracy or scope. For instance, if a paper under review compares methods *A* and *B*, a reviewer might say that it must also compare method *C* in order to be publishable. But simply to complain that more could be done is irrelevant because this is always the case. Rather, the relevant criteria are whether that paper adds to what was known before and whether it has some theoretical interest or practical value. Also, adding method *C* may be a good idea, but the editor might intervene and propose this as a suggestion or recommendation rather than a requirement.

**Fallacious Logic.** Most logical fallacies obtain their apparent plausibility from being subtle variations on other arguments that are valid. Logical fallacies are especially deceptive when their conclusions are already believed or desired.

Fallacies result from invalid variations on the valid argument *modus ponens*: *A*; *A* implies *B*; therefore *B*. The implication "*A* implies *B*" consists of the antecedent *A* and the consequent *B*. Hence, the valid argument *modus ponens* affirms the antecedent. Similarly, the valid argument *modus tollens* denies the consequent: not *B*; *A* implies *B*; therefore not *A*. But other variations are invalid. Affirming the consequent is the logical fallacy: *B*; *A* implies *B*; therefore *A*.

An example is: The plants are yellowish; if plants lack nitrogen, then they become yellowish; therefore the plants lack nitrogen. Likewise, denying the antecedent is also invalid: not *A*; *A* implies *B*; therefore not *B*. A common version of this fallacy is the argument from *missing evidence*. However, an observation of missing evidence *A* has no force in rejecting theory *B* unless it is supplemented with an additional argument showing that evidence *A* would be expected to exist, and perhaps even to be abundant, were theory *B* true. Furthermore, an honest evaluation of theory *B* would also consider whether some other kinds of evidence are relevant and available rather than just eagerly pursuing the easiest possible way to discredit *B*.

Syllogisms have 256 possible forms, of which only 24 are valid. An example of a valid syllogism is: Socrates is a man; all men are mortal; therefore Socrates is mortal. Because most forms are invalid, apart from some training in logic, syllogisms offer numerous opportunities for tricky fallacies.

A *false dilemma* mentions fewer alternatives than actually exist. In the false dilemma "*A* or else *B*; not *A*; therefore *B*," the logical form is valid, but the first premise "*A* or else *B*" is false because of additional possibilities such as *C*. For example, "Either apply nitrogen fertilizer or get yellowish plants" is a false dilemma for many reasons, including the possibilities that a particular soil already has adequate nitrogen without adding fertilizer, or that a virus causes yellowish plants despite adequate fertilizer. Of course, the opposite fallacy also occurs: the "optionitis" of believing that one has more options than reality (or feasibility) actually offers. Some dilemmas are real.

A variant on the false dilemma is the *straw-man argument*. The logical form is this same "*A* or else *B*; not *A*; therefore *B*," where *A* represents an opponent's position and *B* the favored position. However, the premise "not *A*" is supported by attacking the opponent's weakest evidence or a simplistic misrepresentation of the opponent's position. An honest refutation of the opponent's position must instead represent *A* accurately and tackle its strongest evidence and arguments.

Yet another variant on the false dilemma is the *argumentum ad ignorantiam* (argument from ignorance). This fallacy attempts to drive opponents to accept an argument unless they can find a better argument to the contrary. For example, an environmentalist might say "We cannot prove that this pesticide is safe, so we must assume that it is dangerous and outlaw its use." There may or may not be some other good arguments against this pesticide's safety, but an argument from ignorance is not a good reason. The implicit dilemma in an appeal to ignorance is "Give me a better argument, or else accept my argument." But the unmentioned third option is to admit current inability to construct a better argument while still either rejecting the offered argument or suspending judgment.

**Failure of Will.** Given the dishonorable nature of failure of will, this fourth and final category of fallacies is best discussed by adopting Pirie's guise as a

naughty sophist. After all, adroit evasion of knowledge, while still giving every appearance of energetic pursuit of knowledge, requires considerable skill!

Three fallacies are useful to conceal failure of will: privileged cynicism, secret alliance, and personal exemption. Admittedly, these are imperfect means for putting a pretty face on failure of will. But these three fallacies work as well as can be expected, given the inherent challenges of this naughty business. Their principal merits are that these fallacies are unrivaled in their resistance to remediation, and sometimes they can even achieve self-deception!

An effective fallacy for implementing failure of will is *privileged cynicism*: When there is a spectrum of positive to negative opinions about something's merit, the most negative, skeptical, cynical opinion is privileged by being presented and perceived as automatically the view of the sophisticated elite – unlike the naïve and despised view of the ignorant commoners. For instance, the commoners (including most practicing scientists!) may think that much knowledge is readily attained and perfectly solid, whereas presumably the academic elite is steeped in postmodern rejection of knowledge claims, so privileged cynicism declares that *automatically* the latter group has the more sophisticated view. A skilled professor can wield this fallacy to encourage students in a cynical attitude that then becomes the students' passport into the alluring world of the cultural elite. Inside such a culture, cynicism equals sophistication.

This fallacy of privileged cynicism applies readily to science. Students can be lured easily into the mighty gratifying feeling that they, being superior to the gullible commoners, are getting the real, dirty story on what science is. The fallacy of privileged cynicism has great appeal to persons who already feel disappointed or disenfranchised in life for any reason.

A huge advantage of privileged cynicism is its ease. A lackluster high school student, let alone a bright college student, can learn five skeptical or cynical remarks in as many minutes. Furthermore, merely two or three pages suffice for a skilled writer to display a cynical view of science in all its glory, which seems to call for automatic assent from any reader wishing to be numbered among the sophisticates who are in the know. By contrast, a satisfactory account of actual scientific method takes work to write and work to read. Hence, the hard-won sophistication of a working scientist cannot possibly compete with the cynical version of "sophistication" in terms of being offered on the cheap.

A second fallacy for implementing failure of will is *secret alliance*. This wonderfully subtle fallacy involves fighting an intense battle not so much for its own intrinsic importance as for its strategic value in defending an ally in a larger war, while that ally receives so little explicit mention as to remain virtually a secret. Thereby, the real motivations for the battle are not obvious, perhaps even to many of the battle's most prominent combatants on both sides.

The main example in the realm of science is the notorious "science wars" reviewed in Chapter 4. The intensity of this intellectual war, augmented by melodramatic and inflammatory rhetoric, is astonishing and perhaps even

mystifying. Why is it so intense? One might suspect that often the underlying motivations have not been expressed in an entirely forthright manner. Indeed, whether a person's intellectual verdict is that the prospects for human knowledge are dim or bright, and whether that person's emotional reaction to this verdict is sad or happy, are two separate matters. Rather than the usual giddy triumph over vanquished truth, why not express instead a crushing sadness over unrelenting ignorance? This love of ignorance and uncertainty demands some explanation!

Occasionally, there are revealing remarks that arouse suspicions of a secret alliance, although that alliance may be subtle enough to operate at an unconscious level. For example, philosopher Brown (1987:230) remarked, "I have offered here one detailed argument for the now familiar thesis that there is no fundamental methodological difference between philosophy and science. . . . [But] it has become progressively clearer that the sciences cannot provide certainty and have no a priori foundation. . . . [Admittedly,] earlier thinkers believed that both science and philosophy provide certain knowledge of necessary truths. We must conclude that neither do. . . . [The] human intellect . . . seems unable to grasp a final truth." So, chastened science has no truth, and now philosophy can enjoy the same!

For another example, science educator Meyling (1997) mentioned one of his high school students who began with a common-sense, realistic view of science, but in the end she accepted her teacher's "fallibilistic-pluralistic model of epistemology" of "existential uncertainty" and "the tentativeness of science." Meyling quoted her saying that "Truth is relative, we have to get used to that, there are only things that are more correct than others, but there is nothing that is absolutely correct. . . . When you think you know the truth, you force others to think and live that way. . . . This is a claim on absoluteness that cannot be justified – by no one and by no theory." Meyling commented that "I believe that this recognition is far more important than the knowledge about a whole set of scientific 'facts'," and he was particularly pleased that his student extended her new skeptical epistemology to the "ethical level." He mentioned a letter he had received, in which "Sir Karl R. Popper was very pleased with this quote." But Meyling's enthusiasm and Sir Karl's praise notwithstanding, some parents may feel that a science classroom is not a fitting place to encourage ethical relativism or skepticism in other persons' children.

The rather popular idea that science is the sole source and guardian of empirical evidence, and hence of all objective and public knowledge, is a mistake that can seemingly justify failure of will in other realms outside science. But this mistake cannot be supported by mainstream science, which maintains the exact opposite: that scientific thinking, with public evidence as its foremost feature, is also applicable beyond science itself in the humanities and everyday life. Nor can it be supported by insistence on methodological naturalism because this is a stipulatory convention within natural science that is inherently inappropriate in many other disciplines that also use empirical and public evidence. Nevertheless,

this mistaken idea of scientism is easily motivated and long sustained by the most potent of fallacies, failure of will. Frankly, for those persons who heartily want empirical evidence to work for technological comforts *and* not to work for worldview inquiries, simplistic arguments – preferably expressed in a mere sentence or two – should provide welcome and adequate reassurance. On the other hand, for other persons who heartily want empirical evidence to work for technological comforts *and* scientific discoveries *and* worldview inquiries, energetic study of mainstream science and mainstream philosophy should prove fruitful. Getting the most knowledge and benefit from empirical and public evidence requires engaging both the sciences and the humanities, in alignment with the appealing AAAS (1990) vision of science as a liberal art participating in an exciting wider world.

A third and final fallacy for implementing failure of will is *personal exemption*. This fallacy involves mastering fallacies for the purpose of dismantling and evading other persons' arguments, while ignoring the responsibility of detecting and correcting one's own fallacies, as if one has a personal exemption from dealing with truth and reality. The following chapters on probability and statistics examine additional fallacies.

## Summary

Logic is the science of correct reasoning and proof. It addresses the relationship between premises and conclusions, including the bearing of evidence on hypotheses. A deductive argument is valid if the truth of its premises entails the truth of its conclusions and is invalid otherwise. Formal deductive logic begins with a language, axioms, and rules and then derives numerous theorems.

As applied in science, deductive logic argues with certainty from an assumed model to particular expected data. By contrast, inductive logic argues with probability from particular actual data to an inferred general model. In its pursuit of realism and truth, scientific thinking alternates deduction, reasoning from mind to world, and induction, reasoning from world to mind.

The first deductive systems to be axiomatized were syllogisms by Aristotle and then geometry by Euclid. Medieval philosopher-scientists advanced deductive logic considerably. Arithmetic was finally axiomatized only just over a century ago by Peano. The modern vision of deduction, which unites all of its branches into a single unified system built on a base of predicate logic, began with stunningly brilliant work by Frege and by Whitehead and Russell.

The formal system for propositional logic presented here has three axioms and one rule. The axioms for predicate logic are about twice as complicated, but the resulting range of theorems that predicate logic can prove is incomparably greater than the range for propositional logic. Peano arithmetic is presented

with nine axioms. Probability is another branch of deductive logic, but that topic is deferred to the next chapter.

Fallacies have received much interest since Aristotle. Fallacies are best understood and categorized after first recalling the key resources of scientific thinking: admissible and relevant evidence, and deductive and inductive logic. Accordingly, three major categories of fallacies are inadmissible evidence, irrelevant evidence, and fallacious logic. The fourth and final category of fallacies reviewed in this chapter involves a personal rather than a procedural defect: failure of will to pursue the truth.

## Study questions

(1) What are the three interrelated differences between deductive and inductive arguments? Is deduction superior to induction, or are they complementary in scientific thinking?
(2) What are the truth-table definitions for the logical operators Nor and Nand? Why are these operators so extremely useful in computer circuits?
(3) What two main sorts of considerations inform axiom choice for any standard version of a deductive system, such as standard logic or standard arithmetic? What are some applications for nonstandard arithmetic and non-Euclidean geometry?
(4) What is the fallacious *argumentum ad lapidem*, the argument to a stone? Can you contrive an alluring example? How does this fallacy relate to science's presuppositions?
(5) Failure of will to pursue the truth can be implemented by various means, including privileged cynicism, secret alliance, and personal exemption. Give an example of each. Might failure of will be a contributing factor in attacks on science's rationality? Explain your answer.

8

# Probability

Suppose the Smiths tell you that they have two children and show you the family photograph in Figure 8.1. One child is plainly a girl, but the other is obscured by being behind a dog so that its gender is not apparent. What is the probability that the other child is also a girl? This probability question might seem quite simple. But the fact that this problem appeared in the pages of *Scientific American* is a hint that it might actually be tricky (Stewart 1996). Later in this chapter, this problem will be solved, but for now, just remember your initial answer for comparison with the correct solution.

Probability is the branch of deductive logic that deals with uncertainty. Logic occupies three chapters in this book on scientific method, with this chapter being the middle one. The previous chapter concerned other branches of deductive logic: propositional logic, predicate logic, and arithmetic. The next chapter concerns inductive logic, also called statistics. Recall that the PEL model identifies three inputs needed to reach any scientific conclusions: presuppositions, evidence, and logic. Accordingly, science needs functional deductive and inductive logic.

Another reason why the study of probability is important is that errors in probability reasoning are among the most common and detrimental of all fallacies. Probability errors prompt physicians to administer suboptimal treatments. Probability errors prompt juries to render wrong verdicts. And probability errors also cost scientists plenty. Correct probability reasoning is important because scientific research and daily life alike are full of unavoidable practical decisions that must be made on the basis of imperfect information and uncertain inferences. And yet, probability reasoning is rather difficult because it involves precise distinctions and complex relationships that are often subtle and sometimes counter-intuitive. Consequently, some basic training in probability theory can confer substantial benefits.

Figure 8.1 Children in a family. The Smith family has two children. One is a girl, but the other is obscured by the family dog. Reasoning with conditional probability can calculate the probability that the hidden child is also a girl. (This drawing by Susan Bonners is reproduced with her kind permission.)

## Probability concepts

There are two primary concepts of probability, one pertaining to events and one to beliefs. An *objective* or *physical* probability expresses the propensity of an event to occur. For example, upon flipping a fair coin, the probability of heads is 0.5 (because there are two possible events or outcomes, namely, heads or tails, and they are equally likely). A *subjective* or *personal* or *epistemic* probability is the degree of belief in a proposition warranted by the evidence. For example, given today's weather forecast, a given person may judge that the belief "It will rain today" has a 90% probability of being true. Of course, personal and physical probabilities are often interrelated, particularly because personal beliefs are often about physical events. Most events and propositions with low probabilities are not or will not be actualized or true, whereas most with high probabilities are or will be actualized or true.

The concept of probability occurs in a variety of different contexts. A single, unified probability theory needs to work in all of probability's diverse

applications. Consider the following eight common-sense usages of the concept of probability:

(1) A fair coin has a probability of 0.5 of heads, and likewise 0.5 of tails; so the probability of tossing two heads in a row is 0.25.
(2) There is a 10% probability of rain tomorrow.
(3) There is a 10% probability of rain tomorrow given the weather forecast.
(4) Fortunately, there is only a 5% probability that her tumor is malignant, but this will not be known for certain until the surgery is done next week.
(5) Smith has a greater probability of winning the election than does Jones.
(6) I believe that there is a 75% probability that she will want to go out for dinner tonight.
(7) I left my umbrella at home today because the forecast called for only a 1% probability of rain.
(8) Among 100 patients in a clinical trial given drug *A*, 83 recovered, whereas among 100 other patients given drug *B*, only 11 recovered; so new patients will have a higher probability of recovery if treated with drug *A*.

All eight examples use the same word, "probability." To a first approximation, the meaning of probability is the same in all of these examples. Furthermore, in informal discourse, other words could be used with essentially the same meaning, such as "chance" or "likelihood." The common-sense meanings of these examples should be entirely clear to everyone. Nevertheless, these eight examples express a variety of distinguishable concepts.

Example 1 is essentially a definition or theoretical description of what is meant by a "fair" coin, that it has equal chances of landing heads or tails, whether or not any actual coin is exactly fair. It then states the deductive implication regarding tossing two heads in succession. Example 2 is solidly empirical, obviously purporting to convey information about the physical world, namely, about the probability of rain. Example 2 also differs from Example 1 in that the first example's event (a coin toss) is a repeatable event, both in theory and in practice, whereas the second example's event (rain in a particular place and day) is a singular, nonrepeatable event. Incidentally, Example 1 expresses its probabilities as numbers within the range of 0 to 1, whereas Example 2 multiplies such values by 100 to yield a percentage, but this cosmetic difference is not particularly significant.

Example 3 is based on Example 2 but adds an explicit statement about the evidence in support of its assertion. Hence, the unconditional probability in Example 2 is a function of only one thing, the event of rain; whereas the conditional probability in Example 3 is a function of two things, the event of rain given the evidence of the weather forecast. Example 4 expresses a 5% probability of malignancy. In fact, however, the tumor is either benign or malignant – it is *not* 5% malignant and 95% benign. Hence, this probability value of 5% must refer to the present state of knowledge, as contrasted with the actual status of the tumor. This interpretation is reinforced by the expectation

that further knowledge after surgery next week will modify this probability, hopefully to a zero probability of malignancy rather than to the dreaded 100%. Hence, the concept of probability must be capable of handling the addition of new evidence to an existing body of old evidence.

Example 5 says that Smith has a better chance of winning the election than does Jones, but it does not specify whether Smith's chance is great (say, more than 0.5) or small. The problem is that no information is offered about the presence or absence of other candidates. Also note that this example uses no numbers to express its probabilities but rather merely expresses a comparative relationship of one probability being "greater" than another. Example 6 is the first that explicitly recognizes the existence and role of the person expressing a probability judgment, "*I* believe that . . . ." Furthermore, this example is subjective – evidently other persons might come up with different estimates. For instance, someone with no knowledge whatsoever of this woman's plans might pick a probability of 50% to represent maximal uncertainty or ignorance. Hence, the concept of probability needs to be able to model ignorance as well as knowledge.

Example 7 shows probability taking a role not only in personal inferences and beliefs but also in personal decisions and actions. It weighs the personal cost or bother of carrying an umbrella against the potential benefit of not getting soaked. Finally, Example 8 uses data on two recovery rates to derive a conclusion about a probability judgment. Its logical progression is therefore the reverse of that in the first example. Example 1 is representative of deductive thinking that begins with a model or theory (about a fair coin) and then derives conclusions regarding expected observations (of two heads in succession). By contrast, Example 8 is representative of inductive thinking that progresses in the reverse direction. It begins with specific actual observations (regarding 83 and 11 recoveries) and then supports a general theory (about two drugs' relative merits). Hence, the concept of probability is used in both deductive and inductive settings.

In review, a satisfactory theory of probability must encompass events and beliefs, theoretical and physical entities, repeatable and singular events, numerical and comparative expressions, unconditional and conditional probabilities, old and new evidence, knowledge and ignorance, inferences and decisions, and deductive and inductive contexts. Probability concepts must be sophisticated enough to handle complex scientific problems and yet be sensible enough to express simple common-sense applications.

## Four requirements

Probability theory progresses by selecting axioms and then deriving theorems. But what do we want this theory to do? What are its intended applications?

What are the requirements for a satisfactory theory? These requirements are best addressed near the outset of this chapter, even before axiom choice in the next section. Four requirements are specified here, largely following the exceptionally wise and practical book on probability by Sir Harold Jeffreys (1983).

(1) General. An adequate theory of probability must provide a general method suitable for all of its intended applications, including the eight examples of probability concepts in the preceding section. As will be explained in the next chapter, the inductive reasoning in statistics requires no additional axioms beyond those for the deductive reasoning in probability – although decision theory does require one additional axiom. Hence, this chapter's choice of axioms is intended to cover both probability and statistics.

(2) Coherent. The theory of probability and statistics must be coherent or self-consistent. It must not be possible to derive contradictory conclusions from the axioms and any given dataset. Furthermore, all axioms must be stated explicitly, and all subsequent theorems must follow from them. The number of axioms must be small in order to minimize the number of apparently arbitrary choices and to give the foundations great simplicity and clarity.

(3) Empirical. Probability conclusions must be dominated by empirical evidence, not by any presuppositions. Accordingly, probability axioms must be applicable to the physical world but not say specific things about it. For example, consider the scientific finding that the probability is 89.28% that a radioactive decay of a potassium-40 atom emits a positron. Probability theory can be used in this context because it contains no presuppositions about this particular probability, thereby leaving conclusions free to be determined by the evidence. Rather, the only legitimate presupposition of science (including probability theory), which is necessary to render empirical evidence admissible, is that the physical world is real and comprehensible, as was explained in Chapter 5.

(4) Practical. Probability theory must be practical, applicable to real experiences and experiments within reach of human endowments and capacities. It must not require impossible experiments or calculations. The theory must provide for occasional revisions of erroneous scientific inferences because some mistakes are inevitable. What is required is not perfection but rather recoverability in the light of better analysis or more data. Furthermore, all theories of deduction, including probability, have been shown to have some limitations and imperfections, given sufficient statistical and philosophical inspection – so this must be accepted as a permanent and irremediable situation. But it is better to have a probability theory that can do only 99.999% of what all scientists and philosophers could want and yet is tidy and robust, rather than going after that last 0.001% with a dauntingly erudite and disgustingly complex theory that would still be short of perfection.

Kolmogorov Axioms for Probability

1. The probability of event $E$ is a non-negative real number $P(E) \geq 0$.
2. The probability of the conjunction of all possibilities $\Omega$ is $P(\Omega) = 1$.
3. For mutually exclusive events $E_1$, $E_2$, …, the probability of this conjunction of events equals the sum of the individual probabilities: $P(E_1 \cup E_2 \cup \ldots) = P(E_1) + P(E_2) + \ldots$.

**Figure 8.2 The three Kolmogorov axioms for probability. They were formulated by Andrey Kolmogorov in 1933.**

## Probability axioms

Some notation is needed to express probability axioms. The probability of event or belief $X$ is denoted by $P(X)$. The negation of $X$ is denoted by $\sim X$. The members of a set are listed in brackets, such as {1, 2} being the set composed of the integers 1 and 2. The union of $X$ and $Y$ is designated by $X \cup Y$, the set containing everything that belongs to $X$ or $Y$ or both. The intersection of $X$ and $Y$ is designated by $X \cap Y$, the set containing everything belonging to both $X$ and $Y$. For example, if $X = \{1, 2\}$ and $Y = \{2, 3, 4\}$, then the union is $X \cup Y = \{1, 2, 3, 4\}$ and the intersection is $X \cap Y = \{2\}$. By definition, two sets are mutually exclusive if they have no members in common, such as {2} and {3, 4}. Also, by definition, sets are jointly exhaustive if there are no other possibilities. The universal set of all possible outcomes is denoted by $\Omega$.

Figure 8.2 gives the Kolmogorov (1933) axioms for probability. Remarkably, from these three simple axioms, as well as the inherited axioms of predicate logic and arithmetic, all probability theorems can be derived.

For instance, if the probability of $X$ is 0.7, what is the probability of $\sim X$? None of these three axioms can provide an answer for this question. However, the required theorem can be derived readily from these axioms in order to calculate the answer. The union $X \cup \sim X$ is a universal set of all possible outcomes $\Omega$, so axiom 2 yields $P(X \cup \sim X) = P(\Omega) = 1$. Because $X$ and $\sim X$ are mutually exclusive, axiom 3 yields $P(X \cup \sim X) = P(X) + P(\sim X)$. Combining these two results gives $P(X) + P(\sim X) = 1$, and finally rearranging terms yields $P(\sim X) = 1 - P(X)$. Hence, if the probability of rain equals 0.7, then the probability of no rain equals 0.3.

A conditional probability is the probability of $X$ given $Y$, or $X$ conditional on $Y$, and is denoted by $P(X \mid Y)$ where the vertical bar "|" means "given." It can be defined in terms of unconditional probabilities: $P(X \mid Y) = P(X \cap Y) / P(Y)$, provided that $P(Y)$ does not equal 0. As a simple example, assume that a class has six girls with blue eyes and four with brown eyes, and five boys with blue eyes and eight with brown eyes. The conditional probability that a student has blue eyes, given that the student is a girl, is $P(\text{blue} \mid \text{girl}) = 6 / (6 + 4)$, or 0.6.

The conditional probability that a student has blue eyes, given that the student is a boy, is P(blue | boy) = 5 / (5 + 8), or about 0.38. For comparison, the unconditional probability that a student has blue eyes is P(blue) = (6 + 5) / (6 + 4 + 5 + 8), or about 0.48.

Conditional probabilities can be surprisingly tricky. Returning to the opening question about the Smith family photograph shown in Figure 8.1: If the Smiths have two children, one of whom is a girl but the other is obscured by the family dog, what is the probability that the other is a girl? Because the obscured child is a specific child, the correct answer is simply 1/2. The same answer applies to any other specific child, such as the youngest child or the oldest child.

However, a different question can have a different answer. Consider instead the question: If the Millers have two children, one of whom is a girl, what is the probability that the other is a girl? Let B and G denote boy and girl, and make the simplifying assumptions that P(B) = P(G) = 1/2 and that gender is an independent factor (although, in fact, boys are slightly more numerous than girls and gender is not independent in the rare case of identical twins – but the topic here is probability theory rather than reproductive biology). Then there are four equally probable gender sequences for the Miller children with the letters arranged in order of birth: BB, BG, GB, and GG. Because we know that the Millers have at least one girl, the sequence BB is eliminated. That leaves three equally likely cases, and in just one of those cases (GG) is the other child also a girl. Hence, the required conditional probability that the other child is a girl is actually 1/3. Therefore, far from being *equally* likely that the other child is a boy or a girl, actually it is *twice* as likely that the other child is a boy.

At this point, you might recall your initial answer to the probability problem about the Smith family photograph, which was posed at the start of this chapter, in order to check whether you had gotten it right. But, even if your answer was correct, was it more than a lucky guess? Was the reasoning behind your answer adequately precise so that you could distinguish and solve both of these probability problems? The crucial difference between these two problems is that the Smith family has two specific children, the visible child and the obscured child, and only the obscured child might be a boy; whereas the Miller family has two unspecific children, so either child might be a boy.

The Kolmogorov axioms are not uniquely suitable for probability because other axiom sets can be chosen that are equivalent and exchangeable, supporting the same probability theorems. For instance, Salmon (1967:59–60) used four axioms expressed with conditional probabilities, the first being that $P(X \mid Y)$ is a single number $0 \leq P(X \mid Y) \leq 1$. These axioms are equally suitable. But some choices may render the proof of a given theorem somewhat harder or easier.

Even as there are nonstandard logics, nonstandard arithmetics, and non-Euclidean geometries, so also there are nonstandard probability theories resulting from unusual axioms. Burks (1977:99–164) gave examples. The standard and nonstandard probability theories are equally internally coherent or consistent, although they contradict each other. A probability theory based on

the Kolmogorov axioms or on any equivalent and exchangeable set of axioms suits the four requirements specified in the previous section, but a nonstandard theory does not.

Probability axioms serve to enforce coherence among a set of probability assignments and to derive certain probabilities from others. But they do little to provide probability assignments. Instead, that job is done by probability rules. The main one is the so-called straight rule of induction. It says that "If $n$ *A*s have been examined and $m$ have been found to be *B*s, then the probability that the next *A* examined will be a *B* is $m / n$" (Earman 2000:22). For instance, if 200 university students are surveyed and 146 report that they got a flu shot, then the probability for another student having gotten this shot is 146/200 or 73%. Of course, one would trust this estimate or prediction more if the survey had a random sample of the students – rather than all athletes, or all women, or all graduate students – because special subgroups might introduce a bias.

Besides enforcing coherence, probability axioms also make probabilities meaningful. Given the Kolmogorov or equivalent axioms, probabilities are scaled in the range 0 to 1, so $P(X) = 0$ means $X$ is impossible, $P(X) = 1$ means $X$ is certain, and $P(X) = 0.5$ means that $X$ and $\sim X$ are equally likely. But without probability axioms, $P(X) = 0.3$ or $P(X) = 817.7$ or whatever would be utterly meaningless, communicating nothing about the probability of $X$. Likewise, given the probability axioms and theorems, $P(X) = 0.7$ has a clear implication for $P(\sim X)$; whereas, without a coherent probability theory, there would be no implication whatsoever. Incidentally, the same sentiments apply to arithmetic. That four apples plus three apples equals seven apples is meaningful given the coherent and meaningful context provided by standard arithmetic axioms and theorems; but, without that context, an isolated arithmetic assertion about seven apples would lack meaning utterly. This larger context may be informal and implicit in common sense or may be formal and explicit in probability theory; but, in either case, coherence is essential for meaning in any kind of deductive or inductive reasoning.

## Bayes's theorem

For millennia, there had been interest in quantifying probabilities for gambling and other applications, but exact mathematical formulations developed relatively recently. Early contributors were Pierre de Fermat (1601–1665), Blaise Pascal (1623–1662), Christiaan Huygens (1629–1695), Jakob Bernoulli (1654–1705), Abraham de Moivre (1667–1754), and Daniel Bernoulli (1700–1782). They explored probability in its deductive setting, reasoning from a given model to expected observations.

The first person to explore probability in its inductive setting, reasoning in the opposite direction from actual observations to the model, was Thomas Bayes (1702–1761), whose seminal paper was published posthumously in Bayes (1763)

by his friend Richard Price. Soon thereafter, Pierre-Simon Laplace (1749–1827) independently discovered Bayes's theorem in 1774 and further developed this inductive reasoning, which was called "inverse probability." But then, in the 1920s, an alternative approach called *frequentist statistics* was developed, but that story is better told in the next chapter. The seminal paper by Bayes (1763) is readily available on the Internet and has also been reproduced by Barnard (1958) and Swinburne (2002:117–149).

A simple form of Bayes's theorem is:

$$\mathrm{P}(A|B) = [\mathrm{P}(B|A) \times \mathrm{P}(A)]/\mathrm{P}(B). \tag{8.1}$$

Each term has a conventional name. $\mathrm{P}(A \mid B)$ is the conditional probability of $A$ given $B$, also called the posterior probability of $A$. $\mathrm{P}(B \mid A)$ is the conditional probability of $B$ given $A$, also called the likelihood. $\mathrm{P}(A)$ is the unconditional probability of $A$, also called the prior probability of $A$. $\mathrm{P}(B)$ is the unconditional or prior probability of $B$.

Bayes's theorem can be derived easily from the definition of conditional probability. Recall the definition from the previous section that $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B) \,/\, \mathrm{P}(B)$, so likewise $\mathrm{P}(B \mid A) = \mathrm{P}(A \cap B) \,/\, \mathrm{P}(A)$. Rearranging and combining these two equations yields $\mathrm{P}(A \mid B) \times \mathrm{P}(B) = \mathrm{P}(A \cap B) = \mathrm{P}(B \mid A) \times \mathrm{P}(A)$. Finally, dividing the left and right sides of this equation by $\mathrm{P}(B)$, provided that $\mathrm{P}(B)$ does not equal 0, yields the simple form of Bayes's theorem stated previously.

The salient feature of Bayes's theorem is that it relates the reverse conditional probabilities $\mathrm{P}(A \mid B)$ and $\mathrm{P}(B \mid A)$. These two quantities always have different meanings and usually have different numerical values, sometimes wildly different.

In an important application, let $H$ denote a hypothesis from some theory or model and $E$ denote some evidence or data. Then, a quantity of the form $\mathrm{P}(H \mid E)$ represents inductive reasoning from given evidence to an inferred hypothesis, whereas the reverse conditional probability $\mathrm{P}(E \mid H)$ represents deductive reasoning from a given hypothesis to expected evidence. Recall that these opposite reasoning directions of deduction and induction were depicted in Figure 7.1.

Bayes's theorem is used by statisticians for many purposes, including estimating quantities and testing hypotheses. A convenient form for testing two competing hypotheses $H_1$ and $H_2$ in light of evidence $E$ is the ratio form:

$$\frac{\mathrm{P}(H_1|E)}{\mathrm{P}(H_2|E)} = \frac{\mathrm{P}(E|H_1)}{\mathrm{P}(E|H_2)} \times \frac{\mathrm{P}(H_1)}{\mathrm{P}(H_2)}. \tag{8.2}$$

This equation may be read as the posterior ratio equals the likelihood ratio times the prior ratio. For example, if initial considerations give hypotheses $H_1$ and $H_2$ a prior ratio of 1/5 favoring $H_2$ and a new experiment gives them a likelihood ratio of 200/1 favoring $H_1$, then the posterior ratio of 40/1 reverses the initial preference to instead favor $H_1$.

This middle term, the likelihood ratio $P(E \mid H_1) / P(E \mid H_2)$, is also called the *Bayes factor*. It constitutes an alternative to posterior probabilities or ratios for reporting the results from a Bayesian analysis of some evidence (Kass and Raftery 1995). A large Bayes factor would favor $H_1$, a small factor would favor $H_2$, and a factor near 1 would be rather uninformative.

Another form of Bayes's theorem for two hypotheses is convenient for solving some probability problems:

$$P(H_1|E) = \frac{P(E|H_1) \times P(H_1)}{[P(E|H_1) \times P(H_1)] + [P(E|H_2) \times P(H_2)]}. \tag{8.3}$$

Note that in order to solve for $P(H_1 \mid E)$ on the left, four quantities must be known, as specified on the right. However, in the special case that $H_1$ and $H_2$ are mutually exclusive and jointly exhaustive, either $P(H_1)$ or $P(H_2)$ suffices for determining both of them because of the trivial relationship that $P(H_1) + P(H_2) = 1$. Hence, to solve for $P(H_1 \mid E)$, the required probabilities are $P(E \mid H_1)$, $P(E \mid H_2)$, and either $P(H_1)$ or $P(H_2)$.

A nice little application of Bayes's theorem from statistician Sir Ronald A. Fisher (1973:18–20) concerns black and brown mice. The gene for black fur (*B*) is dominant over the gene for brown (*b*), so the homozygous *BB* and heterozygous *Bb* are black, whereas the homozygous *bb* is brown. Now suppose that a female, known to be heterozygous *Bb* (because her parents were *BB* and *bb*), is mated with a heterozygous male. From basic Mendelian genetics for a diploid organism such as mice, the expectation for the offspring from this mating between two black heterozygous parents is one black homozygous *BB* to two black heterozygous *Bb* to 1 brown homozygous *bb*. Interest now focuses on one of her black daughters. The competing hypotheses are $H_{BB}$ that this black daughter is homozygous and $H_{Bb}$ that she is heterozygous. From Mendelian genetics, the prior probabilities are $P(H_{BB}) = 1/3$ and $P(H_{Bb}) = 2/3$, so the prior ratio is $P(H_{BB}) / P(H_{Bb}) = 1/2$. Further suppose that this black daughter is mated with a brown male (*bb*) and she has a litter with seven offspring, all black. The likelihoods resulting from this experimental evidence are $P(\text{litter} \mid H_{BB}) = 1$ whereas $P(\text{litter} \mid H_{Bb}) = 1 / 2^7 = 1/128$, so the likelihood ratio or Bayes factor is $P(\text{litter} \mid H_{BB}) / P(\text{litter} \mid H_{Bb}) = 128$ in favor of $H_{BB}$. Finally, multiplying the prior ratio from the background information by the Bayes factor from the experimental data gives the posterior ratio $P(H_{BB} \mid \text{litter}) / P(H_{Bb} \mid \text{litter}) = 128/2 = 64$ that favors $H_{BB}$. The posterior probabilities are $P(H_{BB} \mid \text{litter}) = 64 / (1 + 64) = 64/65$ and $P(H_{Bb} \mid \text{litter}) = 1 / (1 + 64) = 1/65$. This example of Bayesian inference is exceptionally tidy because both the prior information from the pedigree and the experimental evidence from the litter are objective and public.

Furthermore, given the reasonably strong evidence from this single litter, any further evidence from additional litters would probably confirm the initial verdict. Indeed, the probability that this initial conclusion $H_{BB}$ is actually false is 1/65, or only about 1.5%. But, if this experiment was repeated and that mouse

produced several more litters, an exceedingly strong conclusion would result because the weight of the evidence grows exponentially with its amount. For instance, five additional and similar litters would give a posterior ratio (or a Bayes factor) of more than $10^{12}$ in favor of $H_{BB}$, which would render $H_{BB}$ practically certain. On the other hand, if the mouse in question were heterozygous, then the investigation would be easier because a single brown offspring would prove the hypothesis $H_{Bb}$ with certainty without any calculations being needed. This definitive conclusion that $H_{Bb}$ is true would be expected to emerge rather quickly because at least one brown mouse among $N$ offspring is expected with probability $1 - 2^{-N}$.

The advantage of posterior probabilities is that they address most directly the foremost question of scientists and scholars about competing hypotheses, namely, which hypothesis is probably true given the evidence. But the advantage of the Bayes factor is that if the prior probabilities are highly controversial or rather inscrutable, then any persons can compute their own posterior probabilities from the reported Bayes factor and their own personal prior probabilities. Also, in the special though fairly frequent case that the Bayes factor is huge – especially $10^{10}$ or more – the Bayes factor delivers an exceedingly strong verdict on its own without the additional labor of calculating and defending particular values for the prior probabilities.

In typical applications, the evidence $E$ is public and settled, whereas the background information in the prior is personal and controversial (unlike the tidy mouse example with clear information determining the prior probabilities). For instance, the shared evidence $E$ could be from a published clinical trial concerning two medications $X$ and $Y$, whereas the prior information of individual physicians could be their own experiences of success or failure from giving patients those medications, which might vary considerably from physician to physician. The public evidence $E$ needs to be reasonably strong in order to have greater influence than one's personal prior information (which might be contrary to the clinical trial) and thereby to convince most persons, or at least to interest them.

From its start in 1763 to the present day, Bayes's theorem has been applied extensively in wonderfully diverse contexts across the sciences and the humanities. In his introduction to Bayes's paper, Richard Price touted Bayes's contribution as "a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter," and said that "it is also a problem that has never before been solved" (Bayes 1763). Bayes's own examples concerned the positions of balls on a table and the proportions of blanks and prizes in a lottery. But Price's introduction also mentioned the potential application in philosophical or theological arguments "from final causes for the existence of the Deity," for which this new inverse probability is "more directly applicable" than the previous sorts of probability reasoning. Bayes's paper was published in the *Philosophical Transactions* of the Royal Society, the oldest scientific

society in the world, and both the Rev. Bayes and the Rev. Price were fellows of that prestigious society. The wide range of interests and applications that they foresaw for probability in this new context of inductive reasoning were characteristic of the philosopher-scientists of their times. Indeed, the second Charter of 1663 of the Royal Society (which replaced the first Charter of 1662 with greater privileges by which the Society has since been, and continues to be, governed) stated the purpose of "further promoting by the authority of experiments the sciences of natural things and of useful arts, to the glory of God the Creator, and the advantage of the human race." McGrayne (2011) has written a brilliant and entertaining account of how Bayes's theorem has contributed to many important developments, including cracking the German's Enigma code during World War II, discovering how genes are controlled and regulated, and implementing spam filters for email. At present, Bayes's theorem is used extensively in science and technology, as well as in philosophy and the social sciences. Countless consumer goods are what they are today in part because of Bayes's theorem helping scientists to optimize their inferences and decisions during the research and development that has improved those products. Likewise, Bayes's theorem has added considerable clarity to many important conversations in the humanities.

## Probability distributions

A dozen probability distributions are used frequently and dozens more occasionally. A few of the most important ones are mentioned here. A probability distribution specifies the probability $y$ over the range of the variable $x$. The height of a probability curve is adjusted such that the area under the curve is 1, in keeping with the second Kolmogorov axiom that $P(\Omega) = 1$.

The uniform distribution is the simplest one. Over the range $0 \leq x \leq 1$, its probability is $y = 1$ inside this range, and $y = 0$ elsewhere. More generally, over the range $a \leq x \leq b$, the probability is $y = 1 / (b - a)$ inside this range, and $y = 0$ elsewhere. For instance, over the range $-1 \leq x \leq 1$, the probability is $y = 0.5$ inside this range, and $y = 0$ elsewhere.

The normal or Gaussian distribution is the most prominent probability distribution. It has the equation:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8.4}$$

where $\mu$ is the mean, which locates the peak of this familiar bell-shaped curve, and $\sigma^2$ is the variance, which measures the width or spread of the distribution. The square root of the variance, $\sigma$, is called the standard deviation. The standard normal distribution has $\mu = 0$ and $\sigma^2 = 1$. For the standard normal distribution,

from plus to minus 1 standard deviation accounts for about 68.27% of the area under the curve, 2 standard deviations for 95.45%, and 3 for 99.73%.

The central-limit theorem states that under mild conditions, the sum of a large number of random variables is distributed approximately normally. For instance, an easy method for generating random variables with a nearly standard normal distribution is to sum 12 random variables from the uniform distribution with range $0 \leq x \leq 1$ and then subtract 6. Observational errors are often caused by the cumulative effects of a number of uncontrolled factors, giving these errors an approximately normal distribution. For $N$ replicate observations with errors having a standard deviation of $\sigma$, the standard error of the mean of those $N$ replicates equals $\sigma \, / \, N^{0.5}$. The normal distribution has the advantage of being very tractable mathematically.

The log-normal distribution applies to a random variable whose logarithm is normally distributed. Whereas the normal distribution arises from summing a number of random variables, the log-normal distribution arises from multiplying a number of random variables, all of which are positive. An amazing number of things in the physical, biological, and social sciences approximate a log-normal distribution. Examples include the diameter of ice crystals in ice cream or oil drops in mayonnaise, the abundance of species (bacteria, plants, or animals), latency periods for many human diseases, city sizes, and household incomes.

The binomial distribution describes the number of successes in a sequence of $N$ independent trials, each of which yields success with probability $p$ and failure $(1 - p)$. A simple example is tosses of a fair coin, with heads and tails equally probable. As the number of coin tosses becomes large, the binomial distribution approximates the normal distribution.

Another distribution is the Poisson distribution, which expresses the probability of a number of events occurring within a fixed period of time if these events occur with a known average rate and independently of the time since the last event. For instance, it applies to decays of atoms in a sample of a radioactive material.

Many common probability distributions, including the normal, binomial, and Poisson distributions, belong to an important class, the exponential family. Accordingly, mathematical results proven for the exponential family have a wide range of applications. However, for this book on scientific method, this brief section on probability distributions must suffice. Any further study of probability distributions is better left to texts on probability and statistics.

## Permutations and combinations

Many probability problems, especially in the context of games and gambling, involve a number of possible outcomes that are equally probable. Such problems can be solved by counting numbers of permutations or combinations.

For $R$ events or experiments such that the first event has $N_1$ possible outcomes, and for each of those the second event has $N_2$ possible outcomes, and so on up to the $R^{\text{th}}$ with $N_R$ outcomes, there is a total of $N_1 \times N_2 \times \cdots \times N_R$ possible outcomes. For example, how many different license plates could be made with three digits followed by three letters? The solution is

$$10 \times 10 \times 10 \times 26 \times 26 \times 26 = 17{,}576{,}000. \tag{8.5}$$

A permutation is a distinct ordered arrangement of items. For example, for the set of letters *A* and *B* and *C*, all possible permutations are *ABC*, *ACB*, *BAC*, *BCA*, *CAB*, and *CBA* – which number 6, because the first choice has 3 options, the second 2, and the third 1, for a total of $3 \times 2 \times 1 = 6$ permutations. The general rule is that for $N$ entities, there are $N \times (N-1) \times (N-2) \times \cdots \times 3 \times 2 \times 1$ permutations. This number is called "$N$ factorial" and is denoted by $N!$ For example, $1! = 1$, $3! = 6$, and $5! = 120$. Also, by definition $0! = 1$. As a simple probability problem, presuming that these three letters are drawn at random, what is the probability of a drawing starting with the letter *A*? It is 1/3 because there are two permutations satisfying this condition (*ABC* and *ACB*) and there are six permutations all told.

Sometimes the entities are not all unique, as in the set *A*, *A*, *B*, and *C*, which has two letters *A* that are alike. For $N$ objects, of which $N_1$ are alike, $N_2$ are alike, and so on up to $N_R$ alike, there are $N! / (N_1! \times N_2! \times \cdots \times N_R!)$ permutations. Hence, this set of letters has $4! / (2! \times 1! \times 1!) = 24/2 = 12$ permutations. Presuming that these four letters are drawn at random, what is the probability of a drawing starting with both *A*s? It is 1/6 because there are 2 permutations satisfying this condition (*AABC* and *AACB*) and there are 12 permutations all told.

A combination is a particular number for each of several different entities or outcomes for which the order does not matter. For instance, consider selecting three items from the five items *A*, *B*, *C*, *D*, and *E*. There are five ways to select the first item, four for the second, and three for the third, so there are $5 \times 4 \times 3 = 60$ permutations that distinguish different orderings. But every group of three items, such as *A* and *B* and *C*, gets counted $3! = 6$ times, as was explained earlier. So the number of combinations, which do not distinguish different orderings, of three items selected from five is 60/6 or 10 and all of them are equally probable. The general rule is that $N$ objects taken $R$ at a time have $N! / ((N-R)! \times R!)$ possible combinations. Presuming that three letters are drawn from these five at random, what is the probability of drawing a combination that includes the letters *A* and *B*? It is 3/10 because there are 3 combinations satisfying this condition (*ABC*, *ABD*, and *ABE*) and there are 10 combinations all told.

## Probability fallacies

There is a substantial literature on probability fallacies in medicine, law, science, and other fields. An especially common one is the *base rate fallacy*, also called

the *false positive paradox*, which results from neglecting the base rate or prior probabilities. But because that neglect automatically leads to confusion between $P(H|E)$ and $P(E|H)$, an equally suitable name would be the *reversed conditionals fallacy*. Stirzacker (1994:25) proposed an example that I have expressed concisely as follows:

> A simple medical problem involves three facts and one question. The facts are: (1) A rare disease occurs by chance in 1 in every 100,000 persons. (2) If a person has the disease, a fairly reliable blood test correctly diagnoses the disease with probability 0.95. (3) If a person does not have the disease, the test gives a false diagnosis of disease with probability 0.005. If the blood test says that a person has the disease, what is the probability that this is a correct diagnosis?

Most people, including many physicians, answer that the probability of disease is about 95%. However, from plugging the numbers into Bayes's theorem, surprisingly the correct answer is $(0.95 \times 0.00001) / [(0.95 \times 0.00001) + (0.005 \times 0.99999)]$, or only about 0.2%. This is *drastically* different from 95%, and it strongly supports the exact opposite conclusion! As Stirzaker (1994:26) remarked, "Despite appearing to be a pretty good test, for a disease as rare as this the test is almost useless." For every real instance of the disease detected by that test, there would be more than 500 false positives, so the results could hardly be taken seriously. At best, such a test might offer economical screening before administering another more expensive, definitive test.

What went wrong to give that incorrect answer? Let $H_W$ and $H_S$ be the hypotheses that the person is actually well or sick, and let $E$ be the evidence of a blood test indicating disease. What is given is that $P(H_S) = 0.00001$ and hence $P(H_W) = 0.99999$, that $P(E \mid H_S) = 0.95$, and that $P(E \mid H_W) = 0.005$; and what is required is $P(H_S \mid E)$. The incorrect answer results from ignoring all but one fact, that $P(E \mid H_S) = 0.95$, and assuming erroneously that the reverse conditional probability $P(H_S \mid E)$ has this same value. But ignoring the base rate is a fallacy. Indeed, all three facts given in this problem are needed to obtain the correct solution. That most people in general, and many doctors in particular, make this common blunder in probability reasoning is alarming and potentially dangerous.

Two additional probability fallacies are problematic in the context of law, as emphasized in a seminal paper by Thompson and Schumann (1987): the *Prosecutor's Fallacy* and the *Defense Attorney's Fallacy*. This context is tremendously important because, increasingly, criminal cases have involved scientific evidence and statistical arguments. They gave an example that I have expressed concisely as follows:

> A simple legal example of probability reasoning involves two facts and one question. The facts are: (1) various other kinds of information give a prior probability of 0.1 that the suspect is guilty of committing a murder; (2) a sample of the murderer's blood found at the scene of the crime matches the suspect's rare blood type found in only one person

in 100. The question is: How much weight should be given to this evidence from the laboratory blood test?

Thompson and Schumann presented this probability problem to 73 college undergraduates and asked them to evaluate the following two arguments. The first is an example of the Prosecutor's Fallacy and the second of the Defense Attorney's Fallacy. I have edited their texts slightly.

> The prosecution argued that the blood test is highly relevant. The suspect has the same blood type as the attacker. This blood type is found in only 1% of the population, so there is only a 1% chance that the blood found at the scene came from someone other than the suspect. Since there is only a 1% chance that someone else committed the crime, there is a 99% chance that the suspect is guilty.

> The defense argued that the evidence about blood types has very little relevance. Admittedly only 1% of the population has the rare blood type. But the city where the crime occurred has a population of 200,000 so this blood type would be found in about 2,000 persons. Therefore, the evidence merely shows that the suspect is one out of 2,000 persons in the city who might have committed the crime. A one-in-2,000 chance or 0.05% probability has little relevance for proving that this suspect is guilty.

What do you think of these two arguments? Of the 73 college students, about 29% thought the prosecution's argument is correct and 69% thought the defense's argument is correct (including a few students who judged both arguments correct). In fact, both of these arguments are fallacious.

Let $H_G$ and $H_I$ denote the hypotheses that the suspect is guilty or innocent and $E$ denote the evidence of the matching blood type. We are given that $P(H_G) = 0.1$, so $P(H_I) = 0.9$, and also that $P(E \mid H_I) = 0.01$. The prosecution's argument assumes that $P(H_G \mid E) = 1 - P(E \mid H_I)$. But this is a fallacy because these reversed conditional probabilities are not properly related (in addition to the problem that the prior information is ignored).

On the other hand, the defense's argument adds a new fact, that the population of the city is 200,000. But that fact is irrelevant. This number does not appear when this problem is framed properly by Bayes's theorem. That it is irrelevant can be seen easily by attempting to insert such a number into the previous problem, which is analogous, about a blood test for a rare disease. Whether the patient lives in a large city of 10,000,000 persons or a small town of 350 persons is obviously irrelevant as regards this patient's diagnosis.

What is the probability of guilt given the match, $P(H_G \mid E)$, according to Bayes's theorem? Again, we know $P(H_G)$, $P(H_I)$, and $P(E \mid H_I)$. However, we are not supplied any value for $P(E \mid H_G)$, so not enough information has been given to solve this problem. Nevertheless, for the sake of argument, if we assume that $P(E \mid H_G)$ is quite close to 1, then by Bayes's theorem the answer is

$(0.01 \times 0.9) / [(0.01 \times 0.9) + (1 \times 0.1)] = 0.009/0.109$, or about 8% probability that the suspect is guilty, given the prior information and the blood test. Note that this value is in between the 99% probability of the prosecution and the 0.05% probability of the defense. Also note that the probability of matching blood types $P(E \mid H_I)$ is required by Bayes's theorem in order to calculate the reverse conditional probability $P(H_I \mid E)$, so the defense's argument that the former has little relevance is clearly fallacious.

Thompson and Schumann concluded on a somber note. "The use of mathematical evidence is likely to increase dramatically in the future . . . and legal professionals will increasingly face difficult choices about how to deal with it. Because their choices will turn, in part, on assumptions about the way people respond to mathematical evidence, now is an opportune time for social scientists to begin exploring this issue. Our hope is that social scientists . . . will be able to answer the key underlying behavioral questions so that lawyers and judges may base decisions about mathematical evidence on empirical data rather than unguided intuitions."

Whether an application of probability theory is in medicine or law or science or whatever, such precisely are the choices: reliable inferences based on empirical evidence, or else unreliable inferences based on unguided intuitions. Even a basic education in probability theory, such as this brief chapter provides or at least begins, can reduce probability fallacies.

## Summary

Probability is the propensity for an event to occur or for a proposition to be true. A convenient scaling for probabilities ranges from 0 to 1, with 0 representing certain falsity, 1 representing certain truth, and intermediate values representing uncertainty.

To do business with physical reality, probability theory must meet four basic requirements. It must be general, suitable for all of its intended applications, including deductive probability reasoning and inductive statistics reasoning. It must be coherent. It must not make specific assertions about the physical world in advance of observation and experimentation so that probability conclusions can be dominated by empirical evidence, not by any inappropriate presuppositions. And it must be practical, applicable to real experiences and experiments within reach of human endowments and capacities.

Probability theory is built on predicate logic and arithmetic, requiring the three Kolmogorov probability axioms (or an equivalent set of axioms). Probability axioms serve to enforce coherence among a set of probability assignments and to make probabilities meaningful. The definition of conditional probability is $P(X \mid Y) = P(X \cap Y) / P(Y)$, provided that $P(Y)$ does not equal 0. In addition to

axioms, probability theory also needs rules to assign probabilities, the principal one being the straight rule of induction. And, for certain problems, counting numbers of permutations and combinations of equally probable events can provide probability assignments.

Bayes's theorem is easily derived from the definition of conditional probability. The salient feature of Bayes's theorem is that it relates the reverse conditional probabilities $P(A \mid B)$ and $P(B \mid A)$. A simple form of Bayes's theorem is: $P(A \mid B) = [P(B \mid A)\ (\ P(A)] / P(B)$. Two additional forms are given that are useful for solving various problems. For hypotheses $H_1$ and $H_2$ and evidence $E$, the most common report from a Bayesian hypothesis test is the posterior probabilities $P(H_1 \mid E)$ and $P(H_2 \mid E)$. But an alternative report that is sometimes used is the Bayes factor $P(E \mid H_1) / P(E \mid H_2)$. Since its publication in 1763, Bayes's theorem has had countless applications across the sciences and the humanities.

Several common probability distributions are defined, including the uniform, normal, and log-normal distributions. The normal distribution is particularly important because the central-limit theorem states that under mild conditions, the sum of a large number of random variables is distributed approximately normally. Observational errors often follow an approximately normal distribution.

Probability fallacies are rampant in medicine, law, science, and other fields. An especially common one is the base rate fallacy that results from neglecting the base rate or prior probabilities, which then leads to confusion between the reverse conditional probabilities $P(H \mid E)$ and $P(E \mid H)$, where $H$ denotes some hypothesis and $E$ denotes some evidence. Additional probability fallacies that occur in the context of law are called the Prosecutor's Fallacy and the Defense Attorney's Fallacy. A basic understanding of probability theory helps one make reliable inferences based on empirical evidence rather than unreliable inferences based on unguided intuitions.

## Study questions

(1) The Bakers have three children, of whom two are boys. What is the probability that their other child is a girl?
(2) What four requirements underlie a choice of probability axioms? List the three Kolmogorov axioms. What roles do these axioms serve?
(3) State Bayes's theorem in the ratio form. Give names to the three ratios, as well as a basic explanation of what each term means. What are the differences between the reverse conditional probabilities $P(H \mid D)$ and $P(D \mid H)$ as regards both meaning and numerical value?
(4) What does the central-limit theorem state about the normal distribution? Why do experimental errors often approximate the normal distribution?

(5) A rare disease affects 170 in every 100,000 persons. If a person has this disease, a diagnostic test detects it with probability 0.85; whereas if a person does not have this disease, a false positive occurs with probability 0.003. If the test indicates that a person has this disease, what is the probability that this is a correct diagnosis?

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

Marble Experiment: Problem

**Setup**
Flip a fair coin.
If heads, place in an urn 1 white and 3 blue marbles.
If tails, place in an urn 3 white and 1 blue marbles.

**Hypotheses**
$H_B$: 1 white and 3 blue marbles (WBBB).
$H_W$: 3 white and 1 blue marbles (WWWB).

**Purpose**
To determine which hypothesis, $H_B$ or $H_W$, is probably true.

**Experiment**
Mix the marbles, draw a marble, observe its color, and replace it, repeating this procedure as necessary.

**Stopping Rule**
Stop when a hypothesis reaches a posterior probability of 0.999.

**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 $\pm 7$ 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 $P(H_B)$ and $P(H_W)$ 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 $P(H_B) = 0.1$ 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 $P(H_B) = 0.001$ 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, $P(H_B) = P(H_W) = 0.5$, 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 $P(E \mid H_B) \,/\, P(E \mid H_W)$ 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

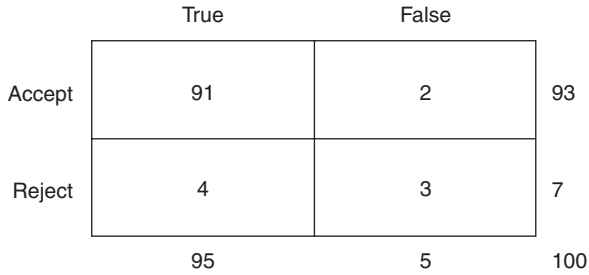


**Figure 9.2 Hypothetical example of error events and rates. A null hypothesis is either true or false, and a test, involving experimental data and a statistical inference, is used to accept or reject the null hypothesis, so there are four possible outcomes, with counts as shown. Also shown are row totals, 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 $\alpha$ is $\mathrm{P}(\text{reject} \mid \text{true}) = 4/95 \approx 0.0421$ and the Type II error rate $\beta$ 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 $n! / (b! \times w!)$ 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 $\alpha$ 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 $\alpha$ 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

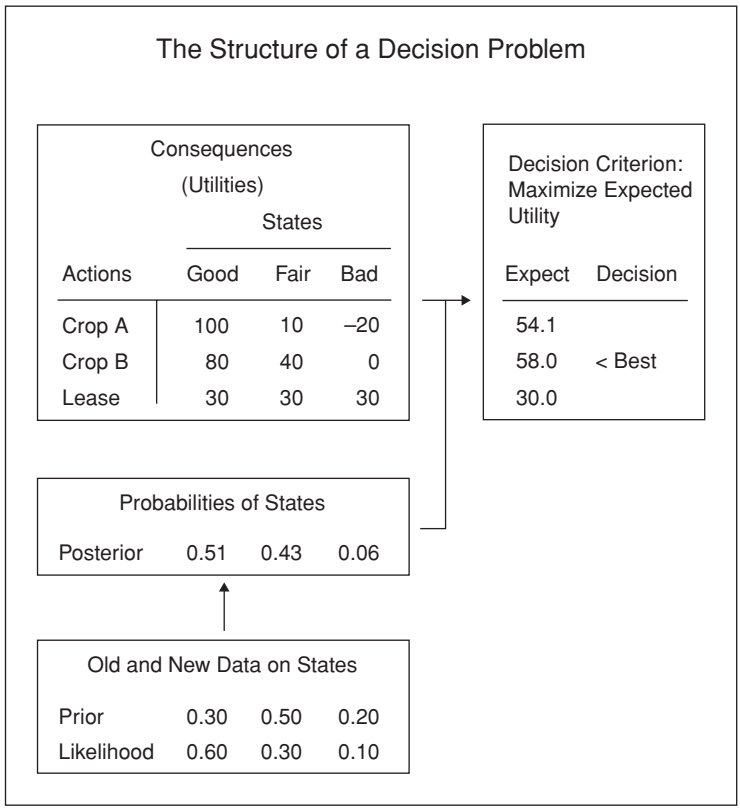


**Figure 9.3 A decision problem about which crop to plant, which concludes that crop B is the best choice.**

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?

10

# Parsimony and efficiency

The principle of parsimony recommends that from among theories fitting the data equally well, scientists choose the simplest theory. It has four common names, also being called the principle of simplicity, the principle of economy, and Ockham's razor (with Ockham sometimes latinized as Occam).

This book's account of science's evidence, which is the "E" portion of the PEL model, takes the form primarily of this chapter's detailed analysis of parsimony. Mere data or observations become evidence when they are brought to bear on hypotheses or theories. This impact of data on theory is guided by several criteria, including the fit of the data with the theory and the parsimony of the theory. Most aspects of evidence are rather obvious to scientists and most evidence is gathered by means of specialized techniques used only within a given discipline. Accordingly, most of what needs to be said about scientific evidence is in the domain of specialized disciplines rather than general principles. However, the one great exception is parsimony, which is not obvious to many scientists, and yet considerations of parsimony pervade all of the sciences, so it is among science's general principles.

Parsimony is not an unusually difficult topic, compared with the ordinary topics routinely studied by scientists. Also, because parsimony pervades all of science, it is easy to find interesting examples and productive applications. Nevertheless, the implementation of parsimony has always faced serious obstacles. In the first place, many scientists seem inclined to think that only a few words, such as "Prefer simpler models," can exhaust the subject. Such complacency does not motivate further study and new insight. Also, the literature on parsimony is scattered in philosophy, statistics, and science, but few scientists read widely in those areas. Yet, each of those disciplines provides distinctive elements that must be combined to achieve a full picture.

In some areas of science and technology, such as in signal processing, the principle of parsimony has already been well understood to great advantage. But, in most areas, a superficial understanding of parsimony has been a serious deficiency of scientific method, costing scientists billions of dollars annually in

wasted resources. Frequently, a parsimonious model that costs a few seconds of computer time can provide insight and increase accuracy as much as would the collection of more data that would cost thousands or millions of dollars. If more scientists really understood parsimony, science and technology would gain considerable momentum.

## Historical perspective on parsimony

Parsimony has been discussed with two distinct but related meanings. On the one hand, parsimony has been considered a feature of nature, that nature chooses the simplest course. On the other hand, parsimony has been deemed a feature of good theories, that the simplest theory that fits the facts is best. These are ontological and epistemological conceptions, respectively, concerning nature itself and humans' theories about nature.

The venerable law of parsimony, the *lex parsimoniae*, has a long history. Aristotle (384–322 BC) discussed parsimony in his *Posterior Analytics*: "We may assume the superiority *ceteris paribus* [other things being equal] of the demonstration which derives from fewer postulates or hypotheses" (McKeon 1941:150). He used parsimony as an ontological principle in rejecting Plato's Forms. Plato (*c.* 427–347 BC) believed that both the perfect Form of a dog and individual dogs existed, but Aristotle held the more parsimonious view that only individual dogs existed. Hence, even something as elemental as the tendency in Western thought to regard individual physical objects as being thoroughly real derives from an appeal to parsimony. Likewise, in his influential commentary on Aristotle's *Metaphysics*, Averroes (Ibn Rushd, 1126–1198) regarded parsimony as a real feature of nature.

Robert Grosseteste (*c.* 1168–1253), who greatly advanced the use of experimental methods in science, also emphasized parsimony, as here in commenting on Aristotle: "That is better and more valuable which requires fewer, other circumstances being equal, just as that demonstration is better, other circumstances being equal, which necessitates the answering of a smaller number of questions for a perfect demonstration or requires a smaller number of suppositions and premisses from which the demonstration proceeds" (Crombie 1962:86). Grosseteste held parsimony not merely as a criterion of good explanations or theories but more fundamentally as a real, objective principle of nature. Thomas Aquinas (*c.* 1225–1274) also espoused a rather ontological version of parsimony, writing that "If a thing can be done adequately by means of one, it is superfluous to do it by means of several; for we observe that nature does not employ two instruments where one suffices" (Hoffmann, Minkin, and Carpenter 1996).

William of Ockham (*c.* 1285–1347) probably is the medieval scholar best known to modern scientists, through the familiar principle of parsimony, often

called Ockham's razor. "It is quite often stated by Ockham in the form: 'Plurality is not to be posited without necessity' (*Pluralitas non est ponenda sine necessitate*), and also, though seldom: 'What can be explained by the assumption of fewer things is vainly explained by the assumption of more things' (*Frustra fit per plura quod potest fieri per pauciora*)" (Boehner 1957:xxi).

Just what does this principle mean? "What Ockham demands in his maxim is that everyone who makes a statement must have a sufficient reason for its truth, 'sufficient reason' being defined as either the observation of a fact, or an immediate logical insight, or divine revelation, or a deduction from these" (Boehner 1957:xxi). However, Ockham's principle of sufficient reason tends to reach modern scientists in a somewhat thinner version of parsimony, merely saying something like: "one should not complicate explanations when simple ones will suffice" (Hoffmann et al. 1996). Ockham insisted that parsimony was an epistemological principle for choosing the best theory, in contrast to his predecessor Robert Grosseteste and his teacher John Duns Scotus, who had interpreted parsimony as also an ontological principle for expecting nature to be simple. In Ockham's view, "This principle of 'sufficient reason' is epistemological or methodological, certainly not an ontological axiom" (Boehner 1957:xxi).

Nicolaus Copernicus (1473–1543) inherited the geocentric cosmology of Aristotle and Ptolemy. It fit the data within observational accuracy, accorded with the common-sense feeling that the earth was unmoving, and enjoyed the authority of Aristotle. However, its one major flaw was lack of parsimony, with its complicated cycles and epicycles for each planet. Consequently, Copernicus offered a new theory: that the earth revolved on its axis daily and journeyed around the sun annually. His main argument featured parsimony: the heliocentric model was simpler, involving fewer epicycles, and the various motions were interlinked in a harmonious system. "I found at length by much and long observation, that if the motions of the other planets were added to the rotation of the earth, and calculated as for the revolution of that planet, not only the phenomena of the others followed from this, but that it so bound together both the order and magnitudes of all the planets and the spheres and the heaven itself that in no single part could one thing be altered without confusion among the other parts and in all the Universe. Hence for this reason . . . I have followed this system" (Dampier 1961:110).

Isaac Newton (1642–1727) further anchored parsimony's importance with the four rules of reasoning in his monumental and influential *Philosophiae Naturalis Principia Mathematica* (Cajori 1947:398–400). Parsimony was the first rule, expressed in a vigorously ontological version concerning nature that echoed words of Aristotle and Duns Scotus: "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances." Newton explained: "To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for

Nature is pleased with simplicity, and affects not the pomp of superfluous causes." Again, parsimony, in a distinctively epistemological version concerning theories about causes, was the second of Newton's rules, corollary to the first: "Therefore to the same natural effects we must, as far as possible, assign the same causes," such as for "respiration in a man and in a beast" and "the reflection of light in the earth, and in the planets." Even his third and fourth rules about experiments and induction rested on the presupposition that nature "is wont to be simple."

Henri Poincaré (1854–1912) related parsimony to generalization: "Let us first observe that any generalization implies, to a certain extent, belief in the unity and simplicity of nature. Today, ideas have changed and, nevertheless, those who do not believe that natural laws have to be simple, are obliged to behave as if it was so. They could not avoid this necessity without rendering impossible all generalization, and consequently all science" (A. Sevin, in Hoffmann et al. 1996). Yet Poincaré also appreciated the subtlety of simplicity, bringing counterpoint with his view that "simplicity is a vague notion" and "everyone calls simple what he finds easy to understand, according to his habits" (A. Sevin, in Hoffmann et al. 1996).

More recently, Albert Einstein (1879–1955) employed parsimony in his discovery of general relativity: "Perhaps the scientist who most clearly understood the necessity for an assumption about the simplicity of [scientific] laws was Albert Einstein. In an informal conversation he once told me about his thoughts in arriving at The General Theory of Relativity. He said that after years of research, he arrived at a particular equation which, on the one hand, explained all known facts and, on the other hand, was considerably simpler than any other equation that explained all these facts. When he reached this point he said to himself that God would not have passed up the opportunity to make nature this simple" (Kemeny 1959:63). Likewise, Einstein spoke of "the grand aim of all science, which is to cover the greatest possible number of empirical facts by logical deductions from the smallest possible number of hypotheses or axioms" (Nash 1963:173). He also remarked that "Everything should be made as simple as possible, but not simpler" (Hoffmann et al. 1996).

Historically, philosophers and scientists have been the scholars who have written about parsimony. More recently, statisticians have also explored this subject, offering two new and important results. First, simple theories tend to make reliable predictions. Second, Bayesian analysis automatically gives simpler theories higher prior probabilities of being true, thereby favoring simpler theories (Jefferys and Berger 1992).

In line with these statisticians, the philosopher Richard Swinburne also saw simplicity as evidence of truth and of reliable predictions: "I seek . . . to show that – other things being equal – the simplest hypothesis proposed as an explanation of phenomena is more likely to be the true one than is any other available

hypothesis, that its predictions are more likely to be true than are those of any other available hypothesis, and that it is an ultimate a priori epistemic principle that simplicity is evidence of truth" (Swinburne 1997:1).

The 1950s was a great decade for parsimony. Statistician Charles Stein published a seminal paper in 1955 that began the literature explaining how parsimonious models can gain accuracy and efficiency. But typical applications require millions or billions of arithmetic steps. For such calculations, reasonably affordable and available digital computers needed transistors, rather than the clumsy vacuum tubes used previously. Physicists John Bardeen, Walter Brattain, and William Shockley co-invented the transistor in 1947 and were awarded the 1956 Nobel Prize in physics. Transistors became increasingly available during the 1950s. Computer programmer John Backus with several associates invented FORTRAN, the first high-level programming language, in 1957. This language made it much easier for statisticians and scientists to use computers. These three resources – new statistical theory, fast transistor circuits, and convenient programming languages – allowed breakthroughs that went far beyond the earlier insights on parsimony from brilliant philosophers such as Aristotle and William of Ockham.

During the subsequent decades, there have been astonishing advances in statistical theory and computing power. Consequently, there are tremendous opportunities to put parsimony to work for gaining accuracy and efficiency, improving predictions and decisions, increasing repeatability, favoring truth, and accelerating progress. Parsimony has had an intriguing history but, more important, it will have an exciting future.

## Preview of basic principles

This chapter's primary means for exploring parsimony are the following three examples of parsimony at work in science. But simplicity is a complicated topic! Accordingly, this section first previews five basic principles.

**Signal and Noise.** Data are imperfect, mixtures of real signal and spurious noise. These terms, "signal" and "noise," originated in the context of radio communication, where a receiver picks up the signal from a transmitter plus noise from various natural and human sources. But, in statistics, these terms are used more generally to refer to treatment or causal effects and random errors. Hence, the data equal the signal plus the noise.

There is a fundamental difference between signal and noise in that the signal ordinarily is relatively simple, caused mostly by only a few major treatment differences or causal factors, whereas the noise typically is extremely complex, caused by numerous small uncontrolled factors. Because the signal is parsimonious, it can be captured or fitted readily by an appropriate parsimonious model, but complex noise inevitably requires a complex model.

**Model Families.** Parsimony is important throughout science, particularly because generalization requires an appeal to parsimony (at least implicitly), as Poincaré emphasized. But parsimony is especially applicable for the common situation in which scientists are considering a model family for analyzing a given dataset. A model family is a sequence of models of the same mathematical form that include more and more parameters. Three such families are used in this chapter.

First, a familiar family is the polynomial family. Let $x$ be the independent variable; $y$ the dependent variable; and $a$, $b$, $c$, and $d$ be constants. Then, the members of the polynomial family are the constant model, $y = a$; the linear model, $y = a + bx$; the quadratic model, $y = a + bx + cx^2$; the cubic model, $y = a + bx + cx^2 + dx^3$; and so on. The data structure that the polynomial model addresses is $N$ paired observations of $x$ and $y$. Given $N$ pairs, the polynomial family has $N$ members with 1 to $N$ terms, the highest-order term being a constant times $x^{N-1}$. The highest member is called the *full or saturated model*. The full model automatically fits the data perfectly, whereas in general the lower-order models fit approximately. For instance, with 7 data points, the highest powers of $x$ are 0, 1, 2, . . . , 6 in the 7 increasingly complex (decreasingly parsimonious) members of the polynomial family.

Second, principal components analysis (PCA) is a common analysis in numerous applications in science and technology, and it also involves a model family. The data structure that PCA addresses is a two-way data table with $R$ rows and $C$ columns. For example, plant breeders often measure yields for $G$ genotypes tested in $E$ environments. PCA provides a suitable model family for such data. A data table with $R$ rows and $C$ columns may be conceptualized geometrically as $R$ points in a $C$-dimensional space, with each point's coordinates specified by the $C$ values in its row (or the reverse, with $C$ points in an $R$-dimensional space). The first principal component is the least-squares line through this high-dimensional cloud of points, meaning that perpendicular projections of these points onto that line maximize the sum of squared distances along that line (and simultaneously minimize the sum of squared distances off that line). The first two principal components specify the least-squares plane in this cloud of points, so they are often graphed to show the structure of the data because they provide the best two-dimensional view of this high-dimensional cloud. Likewise, the first three principal components provide the best three-dimensional approximation to the original cloud, and so on for increasingly complex members of the PCA model family. A common variant of PCA, called doubly-centered PCA, first subtracts the average for each row from each datum in that row, and the same for columns. In agricultural applications, its most common name is the Additive Main effects and Multiplicative Interaction (AMMI) model. For $G$ genotypes and $E$ environments, the highest or full member of the AMMI family has the lesser of $G - 1$ or $E - 1$ principal components. The lowest member, denoted by AMMI-0,

has no principal components but rather only the additive effects, namely, the grand mean, the genotype deviations from the grand mean, and the environment deviations from the grand mean (Gauch 1992:85–96). For instance, the members of the AMMI family for a $7 \times 10$ data matrix are the seven models AMMI-0, AMMI-1, AMMI-2, and so on, up to AMMI-6, which is the full or saturated model that is also denoted by AMMI-F.

Third and finally, multiple linear regression is an extremely popular statistical method that also involves a model family. The data structure that it addresses is a number $M$ of observations or samples for which a dependent variable $\boldsymbol{Y}$ is to be predicted or estimated on the basis of $N$ predictor variables measured for each observation, $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N$, where these variables are bolded to indicate that they are vectors of length $M$. For instance, the data may comprise measurements at several farms of wheat yield, rainfall, soil nitrogen, average August temperature, and altitude, and the objective is to predict wheat yield at each farm from these other four measurements. Multiple linear regression constitutes a model family because there are many choices about which predictors to include and which to exclude in order to get the most accurate predictions. Given $N$ predictors, there are $2^N$ members of the multiple regression model family because each predictor can be in or out. Ordinarily, a rather parsimonious choice will be best. For instance, the best predictor of wheat yield might use only rainfall and soil nitrogen, while discarding the other two variables.

**Statistical Criteria.** Given a model family applied to noisy data, some statistical criterion must be specified to determine the best choice. An important goal is predictive accuracy, but this must be implemented by a specific procedure or equation. Statisticians have devised two basic kinds of strategies.

One strategy involves data resampling techniques, such as cross-validation and the bootstrap. A portion of the data is selected at random, typically about 10% to 25%, and is set aside temporarily to serve as validation data while the model family is fitted to the remaining data. The member with the smallest mean square prediction error for the validation data is selected. Typically, this procedure is repeated many times with different randomizations and the results are averaged for greater accuracy. Often, the final results are based on the selected member of the model family being applied to the recombined, entire dataset.

The other strategy for model choice is to use the Bayesian Information Criterion (BIC), also called the Schwarz Bayesian Criterion (SBC) invented by Schwarz (1978), which approximates the logarithm of the Bayes factor discussed in Chapter 9. The similar Akaike Information Criterion (AIC), or one of the less common alternatives, can also be used. The equations for BIC or AIC contain two terms, one rewarding model fit to the data and another penalizing model complexity (or rewarding model parsimony). Thereby, such criteria strike a balance between fit and parsimony intended to optimize predictive accuracy.

The statistical literature on model choice is extensive and technical (McQuarrie and Tsai 1998; Joo, Wells, and Casella 2010). Data resampling techniques are

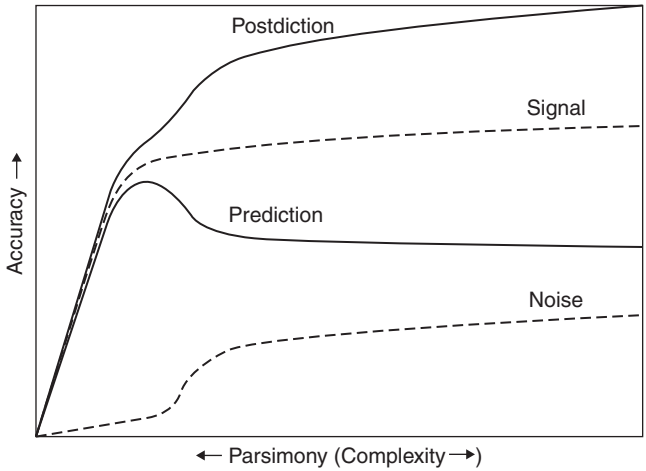


**Figure 10.1 Predictive and postdictive accuracies of models differing in terms of parsimony. The abscissa represents more parsimonious models to the left (and more complex models to the right), and the ordinate shows model accuracy. Imperfect data are mixtures of real signal and spurious noise. Signal is recovered quickly at first as models become more complex, but thereafter signal is recovered slowly. By contrast, noise is recovered slowly at first while much signal is being recovered, then for a brief time noise is recovered more quickly, but thereafter slowly. Postdictive accuracy increases as the signal plus the noise, so it always increases for more and more complex models. But predictive accuracy increases as the signal minus the noise, so it rises to a maximum for some relatively parsimonious model, and thereafter declines.**

rather popular. When several criteria for model selection are compared, often BIC wins. For example, Piepho and Gauch (2001) compared 14 model-selection criteria for simulated genetics data for which the true model was known by construction, and BIC performed best. However, for extremely large models, AIC typically outperforms BIC.

**Ockham's Hill.** Given noisy data analyzed by a model family that is evaluated by a statistical criterion, Figure 10.1 shows what happens. The abscissa depicts a sequence of increasingly parsimonious models moving toward the left (or increasingly complex models moving toward the right). For example, this could be a polynomial family, with its most simple model (the constant model) at the extreme left; then the more complex linear, quadratic, cubic, and higher models progressing toward the right; and finally its most complex model (the full model) at the extreme right. The full model has as many parameters as the data, and its estimates automatically equal the data exactly (such as a quadratic equation with its three parameters automatically going through three data points exactly). The ordinate shows model accuracy or goodness of fit.

Consider first the dashed lines for signal and noise. Because only a few main causal factors determine most of the signal, the relatively simple signal is recovered quickly in early-model parameters and then slowly thereafter. But the response for noise is more involved, with recovery initially slow, then briefly rapid, and then again slow. The initial focus on signal suppresses recovery of noise at first. But after most of the signal has been captured, the focus then shifts to the noise and chance correlations in the noise can be exploited briefly by statistical analyses to accelerate the recovery of noise. Then, after that opportunity has been largely exhausted, noise is recovered slowly.

The data are usually a limited sample from a larger population of interest, such as several hundred persons in a clinical trial who are afflicted by a disease that strikes millions. This distinction between a sample and a population leads to the further distinction between the goal of accurately fitting just the sample data, termed *postdiction*, and the goal of accurately fitting the entire population from which the sample was drawn, termed *prediction*. Nearly always, a scientist's objective is prediction rather than merely postdiction. This distinction between prediction and postdiction is shown by the solid lines in Figure 10.1. It has subtler and greater implications than many scientists realize.

On the one hand, the goal in postdiction is to model or fit the sample data, with no serious concern about a larger population or about the distinction between signal and noise. Recovery of signal and recovery of noise are rewarded alike. Accordingly, the line for postdictive accuracy is depicted as the signal line plus the noise line. The full model at the extreme right automatically recovers all of the signal and noise.

On the other hand, the goal in prediction is to model the entire population of interest. Recovery of signal is rewarded, whereas recovery of noise is penalized because noise is idiosyncratic and has no predictive value. Accordingly, the line for predictive accuracy is depicted as the signal line minus the noise line. Note that the lines for postdiction and prediction are different because of noise. Were the noise negligible, these two lines would be the same.

Quite importantly, these two lines have different shapes, reaching their peaks of maximum accuracy at different places. Postdictive accuracy is automatically maximized by the most complex, full model at the extreme right. But predictive accuracy is maximized for some relatively parsimonious model closer to the left, rather than at the extreme right where the full model equals the data. This means that parsimonious models can be more predictively accurate than their data! That is the principal message of this chapter. The shape of this response for predictive accuracy was given the apt name "Ockham's hill" by MacKay (1992), in honor of William of Ockham.

**Designed Experiments.** Scientific experiments generally have two designs: a treatment design and an experimental design (Gauch 2006). The treatment design specifies the deliberately controlled factors of scientific interest, such as different environments or different genotypes in an agricultural trial. The

experimental design specifies how the treatments are allocated to the experimental units, which usually involves randomization and replication to reduce bias and increase accuracy.

For example, a yield trial could test $G$ genotypes in $E$ environments using $R$ replications, for a total of $GER$ observations. The two-way factorial of $G$ genotypes by $E$ environments constitutes the treatment design, whereas the $R$ replications are involved in the experimental design. Ordinarily, the replications are organized in some specific scheme, such as subdividing the field used for an agricultural trial into a number of subunits or blocks that are smaller and more compact than the field as a whole and, hence, hopefully are rather uniform.

Blocks can be complete, including all treatments; or incomplete, including only some. Importantly, statistical analysis of an incomplete block design pursues two purposes, reducing the estimated errors to increase statistical significance, and adjusting treatment estimates closer to their true values. But complete blocks are less helpful, pursuing only the first of those two purposes.

However, for the present chapter on parsimony, the important message is simply that experiments have *two* designs, the treatment design and the experimental design, and *both* can provide opportunities to gain accuracy. Most scientists are abundantly aware of the value of replication to gain accuracy, although many would not realize that incomplete blocks are more aggressive than complete blocks by virtue of adjusting estimates closer to their true values. But precious few scientists in many disciplines, including agriculture and medicine, are aware of the *other* opportunity to gain accuracy by parsimonious modeling of the treatment design for many common designs. That other opportunity is what this chapter is about. Neglect of this other opportunity is regrettable because its potential for accuracy gain is often several times greater than the potential from replicating and blocking. It is ironic how often scientists implement the smaller of these opportunities to gain accuracy, when the larger opportunity is available but neglected. Of course, best practices require exploiting both opportunities.

## Curve fitting

The first of this chapter's three examples of parsimony is curve fitting using the polynomial model family. The salient features of this example are that the true model is already known exactly, and the noise is also known exactly. Knowing both the signal and noise exactly allows for an unusually penetrating analysis, elucidating principles that subsequently can be recognized in more complex and realistic settings. Obviously, to get an example with signal and noise known exactly requires that we place ourselves in a very unusual position. Such an example must be constructed by us, not offered to us by nature. Accordingly, it must come from mathematics, not from the empirical sciences.

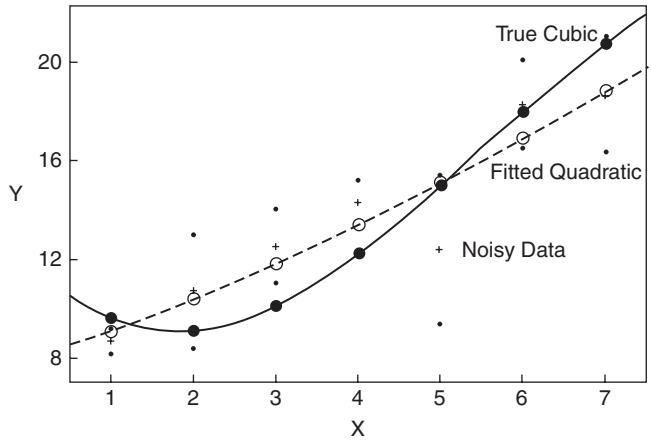


**Figure 10.2 A cubic equation (solid line) is modeled with a quadratic equation (dashed line). Values of the true cubic equation are shown at seven levels of *X* (●). Noisy data are generated at each level for two replicates (•) and their average (+), with these averages having an S/N ratio of 5. Also shown are values for the quadratic equation fitted to the noisy data (○). Note that at every level except the sixth, the fitted quadratic's values are closer to the true values than are the data, the averages over replications. Remarkably, this parsimonious model is more accurate than its noisy data. It achieves a substantial statistical efficiency of 2.10, meaning that on average the quadratic model based on only two replications is slightly more accurate than averages based on twice as much data, four replications. So modeling helps predictive accuracy as much as would collecting twice as much data, but modeling is far more cost-effective when the data are expensive. (Adapted from Gauch 1993 and reproduced with kind permission from *American Scientist*.)**

Figure 10.2 shows a cubic equation, $y = 12.00 - 3.50x + 1.17x^2 - 0.07x^3$, and its values at seven levels, $x = 1, 2, \ldots, 7$. By construction, this cubic equation is the true model or signal, known exactly. To mimic imperfect experimental data, random noise is added that is also known exactly. This noise has a normal distribution adjusted to have a variance of 0.2 times that of the cubic equation's data, which constitutes the signal. By definition, the signal-to-noise (S/N) ratio is the ratio of these variances, namely, $1/0.2 = 5$ in this instance. Frequently, experiments are replicated, which is represented here by showing these noisy data as averages of two replicates (that have twice as much variance as do their averages). Finally, this figure also shows the least-squares quadratic equation fitted to these noisy data, $y = 7.95 + 1.13x + 0.06x^2$.

Note that at every level except the sixth, the fitted quadratic's values are closer to the true values than are the data, the averages over replications. Some persons may find this outcome surprising but, indeed, this model is more accurate than

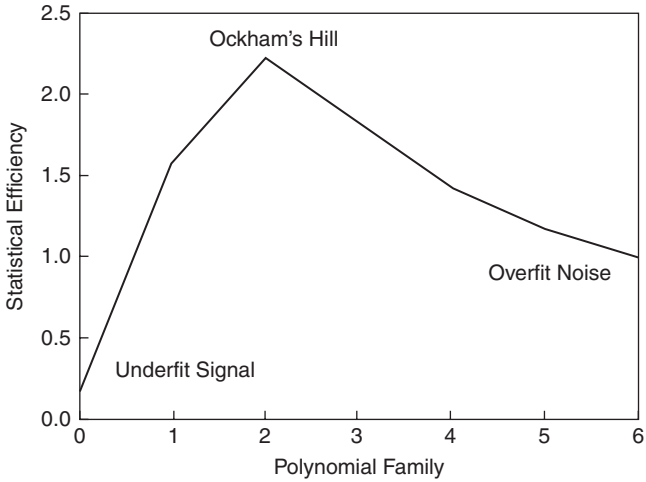


**Figure 10.3 Ockham's hill for the noisy cubic data having an S/N ratio of 5, using the polynomial family encompassing the constant model up to the to sixth-order model. The quadratic model is at the peak of Ockham's hill, achieving the greatest statistical efficiency of 2.22. To the left of the peak, excessively simple models underfit real signal; to the right, excessively complex models overfit spurious noise.**

its data, even though this fitted quadratic model is not the true cubic model! The sum of squares (SS) of differences between the data and true values is 24.67, and the SS of differences between the quadratic model and true values is 11.75. By definition, the statistical efficiency is the ratio of these values, $24.67/11.75 \approx 2.10$. A statistical efficiency of 1 means that a model has the same predictive accuracy as the data, whereas a statistical efficiency above or below 1 implies better or worse accuracy. Because the full model's estimates equal the actual data, its statistical efficiency is automatically exactly 1. Also, a statistical efficiency of 2 or 3 means that a model achieves the same accuracy as would the full model's estimates (namely, averages over replications) based on twice or thrice as many replications. Because this experiment has two replications, the quadratic model is as accurate as would be averages based on $2 \times 2.10 = 4.20$ replications. So, modeling increases accuracy as much as would collecting twice as much data!

The case shown in Figure 10.2 invites three generalizations. Instead of measuring performance with just one set of random noise values, what would be the performance averaged over numerous repetitions with different noise? Instead of presenting results for just the quadratic model, what would be the results for the entire polynomial family? And what would happen at various S/N levels?

Figure 10.3 shows statistical efficiencies for the entire polynomial family for noisy cubic data with an S/N ratio of 5. There are seven data points, so the

polynomial family encompasses the constant model (marked 0 on the abscissa), linear model (1), quadratic model (2), and so on, up to the sixth-order model (6). The statistical efficiency of the quadratic model for the single case analyzed in Figure 10.2 was 2.10, but this figure shows that the average over numerous repetitions with different noise is slightly different, 2.22. Figure 10.3 shows the typical response, Ockham's hill, which was previewed earlier in the line for prediction in Figure 10.1.

The most predictively accurate member of the polynomial family for these noisy cubic data is the quadratic model, achieving a substantial statistical efficiency of 2.22. Efficiency declines in either direction away from the peak, but for different reasons. To the left of the peak, excessively simple models are inaccurate because they underfit real signal. To the right of the peak, excessively complex models are inaccurate because they overfit spurious noise. Optimal accuracy requires a balance between these opposing problems.

Figure 10.4 further generalizes the results for a wide range of noise levels, S/N ratios of 0.1 to 100. Beginning with familiar material from Figure 10.3, note the same results for an S/N ratio of 5 (located about seven-tenths of the way from 1 to 10 on this logarithmic abscissa), with the quadratic model most accurate with its statistical efficiency of 2.22. This figure shows that for rather accurate data having S/N ratios above 16.6, a cubic model is most predictively accurate, achieving a statistical efficiency of 1.82. But because noise increases moving to the left, progressively simpler models are best. However, the fourth-order and higher models never win, including the sixth-order model, which is the full model, equaling the actual noisy data. So which model is most predictively accurate depends on the noise level. It makes sense that as noise increases, fewer of the true model's parameters can be estimated accurately enough to be helpful, until finally only the grand mean, which is the parameter used by the constant model, can resist the onslaught of noise. On the other hand, cleaner data can support more parameters.

Often scientists encounter the entire Ockham's hill, as in Figure 10.3. But Figure 10.4 implies that it is possible to see only the left or right side of the hill. Extremely noisy data make the simplest model win, so there is a monotonic decrease in accuracy for increasingly complex models, thereby showing only the right side of Ockham's hill. Likewise, extremely accurate data make the most complex model win, so only the left side of Ockham's hill is seen.

Depending on the statistic used to express predictive accuracy, Ockham's hill may be inverted, resulting in Ockham's valley instead. Figure 10.3 shows statistical efficiency, which increases with *greater* accuracy; whereas a statistic such as the mean square prediction error increases with *worse* accuracy, so the result is Ockham's valley.

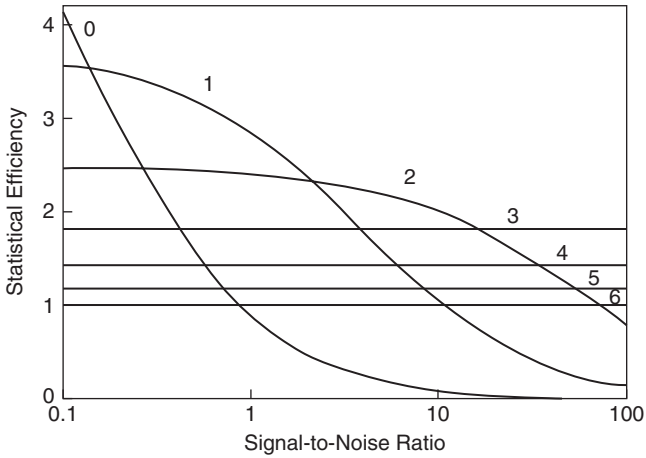


**Figure 10.4** Statistical efficiency of the polynomial family over a range of S/N ratios, with the constant model (0), linear model (1), and others up to the final sixth-order polynomial (6). The constant model is most predictively accurate for extremely noisy data, with S/N below 0.15; the linear model is best for S/N from 0.15 to 2.0; the quadratic model is superior for S/N from 2.0 to 16.6; and the cubic model wins for relatively accurate data, with S/N above 16.6. With seven data points, the full model is the sixth-order equation. This full model is always most postdictively accurate but never most predictively accurate. Notice that for accurate data with an S/N ratio above 16.6, the true cubic model is most predictively accurate; but for noisier data, progressively simpler models are most accurate. Consequently, diagnosing the most predictively accurate member of a model family and determining the true model are distinguishable goals, sometimes having different answers. (Adapted from Gauch 1993 and reproduced with kind permission from *American Scientist*.)

## Crop yields

The second example of parsimony at work is familiar to me from my own research from 1988 to 2012, agricultural yield trials using the AMMI model family. Plant breeders use yield trials to select superior genotypes, and agronomists use them to recommend varieties, fertilizers, and pesticides to farmers. Worldwide, several billion dollars are spent annually on yield trials. These experiments have helped plant breeders to increase crop yields, typically by about 1% to 1.5% per year for open-pollinated crops such as corn and 0.5% to 1% per year for self-pollinated crops such as soybeans. However, there is substantial and worrisome evidence that wheat and rice yield increases have slackened lately to about 0.5% per year, which is considerably less than during 1960 to 1990.

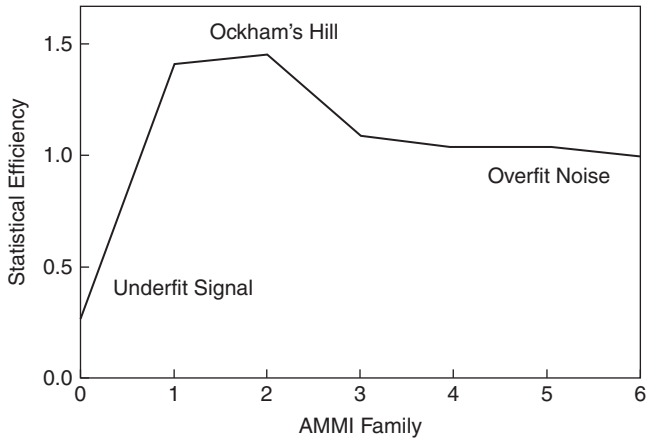


**Figure 10.5 Ockham's hill for the soybean data using the AMMI family. The most predictively accurate member of the AMMI family is AMMI-2, achieving a statistical efficiency of 1.45, but AMMI-1 is almost as accurate.**

The most common type of yield trial tests a number of genotypes in a number of environments that are location-year combinations, often with replication. The dataset used here is a New York State soybean trial having 7 genotypes in 10 environments with 4 replications (Gauch 1992:56). Recall that Figure 1.6 showed a photograph from this soybean yield trial.

The salient feature of this agricultural example is that neither the signal nor the noise is known exactly, quite in contrast to the easy first example with a known cubic equation and known added noise. We receive the soybean data from nature, with the signal and noise already mixed together. Indeed, these data are rather noisy, with typical errors of plus or minus 15%, so they carry only one significant digit. Because the true signal and added noise are not known separately and exactly, the method for calculating statistical efficiencies in Figure 10.5 is more complicated than for Figure 10.3, as explained in detail elsewhere (Gauch 1992:134–153). Greater accuracy for yield estimates also helps in the search for genes affecting yield (Gauch et al. 2011).

Figure 10.5 shows statistical efficiencies for the soybean data using the AMMI family. The AMMI-2 model is at the peak of Ockham's hill, achieving a statistical efficiency of 1.44, with AMMI-1 a close second. To the left of the peak, excessively simple models underfit real signal; to the right, excessively complex models overfit spurious noise.

Looking at the past, between the 1940s and the late 1970s, the principal achievement of the Green Revolution has been to increase dramatically the yields of the major grain crops – wheat, corn, and rice – in favorable environments. This achievement was crucial and it still is, accounting for a large portion of

the world's food supply. But, unfortunately, there were two unintended and detrimental side effects. First, the increased productivity and profitability of major grains prompted many farmers to grow less of other grains, vegetables, and fruits, thereby restricting diets and causing deficiencies in essential vitamins and minerals, especially vitamin A, iron, and zinc. Second, the considerable neglect of marginal environments, which often require different genotypes than favorable environments, meant little benefit for the millions of persons whose food comes mainly from those poor environments.

Looking toward the future, the world's population is currently increasing about 1.2% per year. But the welcome rise out of poverty for millions, especially in China and India, implies greater demand for meat production, which uses much grain. So, crop yields need to increase somewhat more rapidly, about 1.4% per year. The good news is that plant breeders are increasingly poised to address those two deficiencies of the original Green Revolution. For numerous vegetable and fruit species, powerful new genetic tools can allow relatively small projects in several years to increase yields as much as the larger and longer projects of the Green Revolution decades ago. So, those other crops can now become more profitable and contribute to a diverse and wholesome diet, especially because breeders are also selecting for enhanced nutritional traits. Likewise, relatively small projects can more quickly address marginal environments, benefiting many of the world's poorest communities. But the bad news is that yield advances are needed despite trends in many places toward less farm land, less water, and more disease pressure – and despite challenging goals of better environmental stewardship. On balance, a sustained yield increase of 1.4% per year for the next decade or two seems attainable, until projections of world population largely level off. However, there is little margin for agricultural research to be inefficient. If a greater fraction of agricultural researchers learned how to put parsimony to work in order to get the most out of their experiments and data, numerous projects, involving many crops in many nations, would accelerate markedly.

## Crime rates

The third and final example of parsimony is a sociological study of crime rates using the multiple regression model family. This study was first published by Ehrlich (1973). The present account is based on the reanalysis by Raftery (1995), using both frequentist and Bayesian methods.

Raftery explained that "Most sociological studies are observational and aim to infer causal relationships between a dependent variable and independent variables of interest." The choice of predictor variables to measure is guided by theory or background knowledge. But "theory is often weak and vague," so the usual strategy is to play it safe by including a long "laundry list" of candidate

predictors in hopes of not missing anything important. Multiple regression provides a standard statistical method for deciding which candidates to include in the final model. Discarding the irrelevant candidates is important because including useless variables in the model degrades the results for parameters of genuine interest. This setup of many candidate predictors and rather weak theory is quite common, not only in sociology but also in ecology, agriculture, medicine, and many other fields.

This criminological study by Ehrlich was one of the earliest systematic efforts to address the question: Do greater punishments reduce crime rates? As Raftery recounted, there were two competing hypotheses. One hypothesis, which may be denoted $H_1$, is that criminal behavior is "deviant and linked to the offender's presumed exceptional psychological, social, or family circumstances." The competing hypothesis $H_2$ is that "the decision to engage in criminal activity is a rational choice determined by its costs and benefits relative to other (legitimate) opportunities." Ehrlich compiled extensive data on localities from 47 states in the USA. The following paragraph paraphrases Raftery's description of the second hypothesis, inserting in parentheses the numbers for the 15 candidate predictors of crime rates.

> The costs of crime are related to the probability of imprisonment ($X_{14}$) and the average time served in prison ($X_{15}$), which in turn are influenced by police expenditures ($X_4$, $X_5$). The benefits of crime are related to aggregate wealth ($X_{12}$) and income inequality ($X_{13}$) in the surrounding community. The expected net payoff from the alternative of legitimate activities is related to educational level ($X_3$) and the availability of employment, the latter being measured by the unemployment rate ($X_{10}$, $X_{11}$) and labor force participation rate ($X_6$). This payoff is expected to be lower for nonwhites ($X_9$) and for young males ($X_1$). Other possible influences are southern versus northern states ($X_2$), the state population ($X_8$), and the sex ratio or number of males per female ($X_7$). The principal interest is in the probability and length of imprisonment, $X_{14}$ and $X_{15}$, because hypothesis $H_2$ expects greater association with crime rates than does $H_1$.

Beginning with the frequentist analysis, Raftery tried three common methods for choosing statistically significant variables, including the most popular method, stepwise regression. He also tried two variants of Ehrlich's models based on sociological theory. The full model with all 15 predictors was also included as a baseline for comparison. But the results were perplexing. "There are striking differences, indeed conflicts, between the results from different models. Even the statistically chosen models, despite their superficial similarity, lead to conflicting conclusions about the main questions of interest."

Progressing to the Bayesian analysis, Raftery based model selection on the BIC approximation to the logarithm of the Bayes factor. A distinctive feature of the Bayesian approach is that it can incorporate model uncertainty by averaging over multiple models that are all supported well by the data. The 15 candidate predictors imply $2^{15} = 32{,}768$ possible models. Most of these numerous

possibilities are decidedly bad, so calculations can be simplified by averaging over only the reasonably good possibilities using a criterion that Raftery terms Occam's window – although I prefer not to latinize this philosopher's name, so I call it Ockham's window. Models are excluded that (a) are 20 times less likely than the most likely model, corresponding to a BIC difference of 6; and (b) contain predictors for which there is no evidence in the sense that they have more likely sub-models nested within them that omit those predictors. For this study of crime rates, Ockham's window reduced the original 32,768 models to a very manageable 14 models. Raftery reported that this reduction is quite typical. These foremost models within Ockham's window were parsimonious, including only 5 to 8 of the 15 candidate predictors and averaging 6.4 predictors.

The Bayesian analysis showed that the probability of imprisonment ($X_{14}$) has a probability of 98% of having a real effect on crime rates, whereas the length of imprisonment ($X_{15}$) is not particularly significant (only 35%). There is strong evidence that higher crime rates are associated with both educational level ($X_3$) and income inequality ($X_{13}$) with 100% probability, as well as with young males ($X_1$) with 94% and nonwhites ($X_9$) with 83%. But there is no association with crime for aggregate wealth ($X_{12}$), labor force participation rate ($X_6$), and sex ratio ($X_7$). There was also evidence for a negative association between police expenditure ($X_4$, $X_5$) and crime rate, although the causal story for this was not evident.

In these findings, the importance of probability of imprisonment ($X_{14}$), income inequality ($X_{13}$), and educational level ($X_3$) are supportive of hypothesis $H_2$ that emphasizes rational deliberation. On the other hand, sizable effects for young males ($X_1$) and nonwhites ($X_9$) are supportive of $H_1$ that emphasizes social and family influences. These two hypotheses are just different, not mutually exclusive, so there could well be some valid aspects in both.

Regarding statistical paradigm, Raftery detailed numerous advantages of the Bayesian approach over the frequentist approach. Most important, the Bayesian analysis is superior conceptually and operationally because it alone can integrate model selection and parameter estimation. For this particular study of crime rates, a plausible argument can be made that the Bayesian analysis, which favors parsimonious models, provided clearer conclusions.

## Explanation of accuracy gain

How do parsimonious models gain accuracy? First of all, it must be insisted that routinely they do. Cross-validation and related methods prove that for countless applications across science and technology. For example, to cite just one number from just the first of this chapter's examples, at an S/N ratio of 5,

the parsimonious quadratic model achieves an average statistical efficiency of 2.22, and this accuracy gain is absolutely indisputable because the true signal and added noise are known exactly by construction.

Consequently, even if not one person on earth could explain how parsimonious models can be more accurate than their data, this accuracy gain would still stand as an established fact and a great opportunity. However, a fact without an explanation is unsatisfying and sometimes even unconvincing.

There are three interrelated explanations for accuracy gain by parsimonious models. They concern signal–noise selectivity, direct–indirect information, and variance–bias trade-off.

**Signal–Noise Selectivity.** This chapter's preview explained accuracy gain by parsimonious models in terms of signal–noise selectivity. Early model parameters capture mostly the relatively simple signal, whereas late model parameters capture mostly the relatively complex noise, as depicted in Figure 10.1. By selecting the most predictively accurate member of a model family at the peak of Ockham's hill, a signal-rich model is separated from a discarded noise-rich residual.

**Direct–Indirect Information.** Full and parsimonious models make use of the data in strikingly different ways. Recall the Ockham's hill for the soybean data in Figure 10.5. The data for these 7 soybean varieties tested in 10 environments with 4 replications are given in Gauch (1992:56), for a total of 280 yield measurements. For instance, the four replicates for Evans soybeans grown in Aurora, NY, in 1977 were 2,729, 2,747, 2,593, and 2,832 for an average of 2,725 kg/ha.

Suppose that upon checking these numbers against the original field notes, a typographical error is detected: the first replicate should be 2,279 rather than 2,729. What happens to the AMMI-F and AMMI-2 yield estimates upon correcting that error?

The AMMI-F yield estimates equal the actual data, namely, the averages over replicates. Hence, that one value for Evans in Aurora in 1977 changes from 2,725 to 2,613 kg/ha. Nothing else changes.

By contrast, the computation of the AMMI-2 yield estimates involves the entire data matrix. Hence, all yield estimates change, for all 7 varieties in all 10 environments, although most adjustments are rather small.

From the perspective of any given yield, such as that for Evans in Aurora in 1977, there are 4 measurements providing direct information about that yield and 276 measurements providing indirect information about other varieties or other environments or both. The full and parsimonious models are fundamentally different in what they take to be the relevant data. For each and every yield estimate, the full model AMMI-F uses its 4 replicates, whereas the parsimonious model AMMI-2 uses all 280 measurements. That is, AMMI-F uses only the direct information, whereas AMMI-2 uses both the direct and indirect information.

How much does the indirect information help the AMMI-2 yield estimates? There are 4 replicates and the statistical efficiency is 1.44. Hence, AMMI-2 using 4 replicates is as accurate as would be AMMI-F using $4 \times 1.44 = 5.76$ replicates, so the indirect information has helped as much as would adding $5.76 - 4 = 1.76$ more replications. So, the indirect 276 observations equate to 1.76 direct observations, or $276 / 1.76 \approx 157$ indirect observations are as informative as 1 direct observation. The indirect information is dilute, but it is also abundant, and therefore worth incorporating in yield estimates.

Again, how do parsimonious models gain accuracy? The most basic explanation is that they use more data – no magic, no mystery, just more data. This has been understood by statisticians ever since the seminal paper by Stein (1955). This explanation also applies to this chapter's first example, the cubic equation, because the full model uses only the 2 replicates for each of the 7 levels to estimate its values ($y$), whereas a parsimonious model uses all 14 observations to estimate each and every value.

**Variance–Bias Trade-Off.** The third and final interrelated explanation of accuracy gain by parsimonious models concerns a trade-off between variance and bias. Low variance and low bias are both desired. But increasingly complex models in a model family have more variance but less bias, so a trade-off is inevitable. Ockham's hill occurs because a modest amount of both problems is better than a huge amount of either problem. However, this is the most technical of these three explanations, so further details can be relegated to the statistical literature, such as Gauch (1992:134–153).

In review, parsimonious models gain accuracy by retaining early model parameters that selectively capture the relatively simple signal and discarding late-model parameters that selectively capture the relatively complex noise. Most fundamentally, they use available data more aggressively, extracting both direct and indirect information, and they strike an optimal trade-off between problems with variance and bias.

## Philosophical reflection

This chapter's analysis of parsimony has been primarily from a scientific and technological perspective, with special interest in gaining accuracy and efficiency. But greater understanding of parsimony can emerge from adding some philosophical reflection. This section addresses two topics: parsimony and nature, and prediction and truth.

**Parsimony and Nature.** Recall from the historical review that parsimony has two aspects: an epistemological principle, preferring the simplest theory that fits the data, and an ontological principle, expecting nature to be simple. The epistemological aspect of parsimony has been emphasized here because it is part of scientific method. But the ontological aspect also merits attention.

So, is nature simple? For starters, understand that the reality check is itself a *simple* theory about a *simple* world. It declares that "Moving cars are hazardous to pedestrians." This is simple precisely because it applies a single dictum to all persons in all places at all times. The quintessential simplicity of this theory and its world, otherwise easily unnoticed, can be placed in bold relief by giving variants that are not so simple. For example, if nature were more complex than it actually is, more complicated variants could emerge, such as "Moving cars are hazardous to pedestrians, except for women in France on Saturday mornings and wealthy men in India and Colorado when it is raining." Although there is just one simple and sensible formulation of the reality check, obviously there are innumerable complex and ridiculous variants. Regarding cars and pedestrians, a simple world begets a simple theory. Or, to put it the other way around, a simple theory befits a simple world.

Capitalizing on this little example, meager thought and imagination suffice to see parsimony everywhere in the world – in iron atoms that are all iron, in stars that are all stars, in dogs that are all dogs, and so on. Parsimony touches our every thought. But to really understand parsimony, one must move beyond examples to principles.

Induction, uniformity, causality, intelligibility, and other scientific principles all implicate parsimony. Applications of induction to the physical world presuppose parsimony, specifically in the ontological sense expressed strongly by the law of the limited variety of nature. Likewise, the law of causality, that similar causes produce similar effects, is an aspect of simplicity. That nature is intelligible to our feeble human reason shows that some significant features of reality are moderately simple. If nature were not simple, science would lose all of its foundational principles at once.

Yet, the greatest influence of parsimony in scientific method is in the simplicity of the questions asked. Any hypothesis set that expresses a scientific question could in principle always be expanded to include more possibilities, and that action would make sense were the world more complex than it is. Were inductive logic bankrupt, were nature not uniform, were causes not followed by predictable effects, and were nature barely comprehensible, then enormously more hypotheses would merit consideration. Then, science would languish with hopelessly complicated questions that would impose impossible burdens for sufficient evidence. The beginning of science's simplicity is its simple questions.

Having argued that nature is simple, this verdict should not be interpreted simplistically! Indeed, "there is complexity to the whole idea of simplicity" (Nash 1963:182). Simon (1962) offered remarkably keen insights regarding just which aspects of nature scientists expect to be simple and just which aspects they expect to be complex. In essence, the rich complexity of life and ecosystems emerges from the frugal simplicity of basic physical and chemical laws. From

general experience, scientists and engineers ordinarily have a fairly reliable general sense of how simple or complex a given system or problem is.

The verdict on parsimony is that "Ockham's Razor must indubitably be counted among the tried and useful principles of thinking about the facts of this beautiful and terrible world and their underlying causative links" (Hoffmann et al. 1996). Nevertheless, those authors also note the sensible reaction that "the very idea that Ockham's Razor is part of the scientific method seems *strange*... because... science is not about simplicity, but about complexity." The plausible resolution that those authors offer is that simple minds comprehend complex nature by means of ornate models made of simple pieces. The balance between a model's simplicity and the extent to which it approaches completeness requires a delicate and skillful wielding of Ockham's razor. The comments on Hoffmann et al. (1996) by A. Sevin concur: "Our discovery of complexity increases every day.... This good old Ockham's razor remains an indispensable tool for exploring complexity."

**Prediction and Truth.** Predictive success is often taken as evidence of truth. To cite one famous example, using Newton's theory of gravity, Edmond Halley (1656–1742) calculated the orbit of the impressive comet of 1682, which now bears his name, identifying it as the one that had appeared previously in 1531 and 1607, and predicting the time and place of its return in 1759. He did not live to see that return, but it did happen just as he had predicted. His striking predictive success was accepted universally as proof that his theory of comets' orbits was true, or at least very nearly true.

Generalizing from that familiar and yet representative example, predictive success is taken generally as evidence of truth, especially when numerous and diverse predictions are all correct, so that mere luck is an implausible explanation. Indeed, among theories that have attained strong and lasting acceptance among scientists, doubtless one of the most significant and consistent categories of supporting evidence is predictive success.

Nevertheless, this venerable formula, that predictive success implies truth, can be unsettled by interpreting or applying it too simplistically. Indeed, a little reflection on Figure 10.4 should be disturbing, or at least thought-provoking. By construction, a cubic equation is known to be the true model. It is sampled at seven points, with addition of random noise (at an S/N ratio of 5) to mimic measurement errors, and least-squares fits are calculated for the polynomial family. Although a cubic equation is the true model, the cubic model is less predictively accurate than another member of the polynomial family, the quadratic model. So even with the true model entered in the competition, the criterion of predictive accuracy gives the win to a false model! If such problems occur for easy cases with constructed and known models, what happens in the tough world of real scientific research? Does predictive accuracy have no reliable bearing on truth?