SCIENTIFIC METHODS

an online book

Richard D. Jarrard

Dept. of Geology and Geophysics, University of Utah

r.jarrard@utah.edu

Scientific Methods is an online book about the techniques and processes of science and the experience of being a scientist. This book is written by a scientist for scientists. My hope is that it will be browsed by scientists (including graduate students) and read by undergraduates.

Why am I publishing this book online, despite having a willing soft-cover publisher? The main reason is wider availability to readers. A typical science book has a publication run of ~2000 copies, then it goes out of print. Additional factors include educational use and ease of revision. I admit that I would have enjoyed saying that I earned ~25¢/hour by writing this book.

CONTENTS

Chapter 1: Introduction

Overview

Consider the dance of science -- the dance that obsesses us so.

> It's said that in viewing the night sky, the present is illusion. The stars are so distant that I see them as they were millions or billions of years ago, when their light rays began the voyage to my eye. It's said that I am infinitesimally small and transient; the stars will not miss the light my eyes have stolen. They will not notice that they have joined me in the dance.

Technique and style are the framework of dance. Techniques of science are generally the easy part; many are deliberately and systematically taught. For example, throughout our many years of schooling we refine skills such as fact gathering and mathematical analysis. We learn other scientific techniques -- such as statistics, deductive logic, and inductive logic -- in classes that lack the perspective of scientists' needs.

Some techniques are more intangible: critical thinking and analysis, pattern recognition, and troubleshooting of experimental technique. Scientists are not merely technicians; an equally crucial part of the dance is style: how do scientists combine rationality and insight, or skepticism and innovation; how do scientists interact, and what motivates their obsession? These skills seldom are taught explicitly. Instead, they are implicit in the scientific apprenticeship, an excellent but often incomplete educational process.

Who of us has mastered all of the techniques of science? I certainly have not; researching and writing this book have shown me that. Of course, when I recognize that an aspect of my scientific methods is deficient, I am enough of a professional to seek a remedy. More often, I, like Konrad Lorenz's [1962] water-shrew, am not even aware of what is missing:

> The water shrew dashes through its territory at incredible speed, by following the familiar path. "To them, the shortest line is always the accustomed path." Lorenz decided to explore the extent of this habit by removing a stone from a water-shrew's path. When it came racing along, it jumped over the nonexistent stone. It paused in bafflement, backed up and jumped 'over' it again, then finally reconnoitered the anomaly.

How often do we leap missing stones?

* * *

Consider the *science* of science. Let's turn our gaze on our lives, looking beyond the surface interplay of experiment and theory. What are we scientists doing, and what tools are we using?

We've left such introspection to philosophers, but their goals differ from ours. They deal in abstracts: what rules do scientists follow, and how should the process of science change? We scientists generally prefer the more pragmatic approach of just doing, not talking about doing. Are we too busy, or too confident in our established routines, to analyze what we are doing? Why are virtually all of the books on scientific methods written by philosophers of science, rather than by scientists?

"It is inevitable that, in seeking for its greatest unification, science will make itself an object of scientific investigation." [Morris, 1938]

* * *

This book was originally intended as 'How to do science', or 'How to be a scientist', providing guidance for the new scientist, as well as some reminders and tips for experienced researchers. Such a book does not need to be written by the most expert or most famous scientist, but by one who likes to see the rules of play laid out concisely. It does need to be written by a working scientist, not by a philosopher of science. The first half of the book, called 'Scientist's Toolbox', retains this original focus on what Jerome Brumer called the structure of science -- its methodologies and logic.

This objective is still present in the second half of the book, 'Living Science'. In researching that section, however, I was fascinated by the perspectives of fellow scientists on 'What it is like to be a scientist.' Encountering their insights into the humanity of science, I found resonance with my already intense enjoyment of the process of science. Gaither and Cavazon-Gaither [2000] provide many additional scientific quotations on the experience of science.

* * *

Consider the *process* of science.

Knowledge is the goal of science: basic research seeks reliable knowledge, and applied research seeks useful knowledge. But if knowledge were our primary goal *as scientists*, we would spend much of our available time in reading the literature rather than in slowly gathering new data. Science is not static knowledge; it is a dynamic process of exploring the world and seeking to obtain a trustworthy understanding of it. Everyone practices this process, to some extent. Science is not the opposite of intuition, but a way of employing reality testing to harness intuition effectively and productively.

As we explore the scientific process in this book, we will attempt to answer some of the following questions.

- History: What are the *essential* elements of scientific method?
- Variables: How can I extract the most information from my data?
- Induction and pattern recognition: If I cannot think of an experiment to solve my problem, how can I transpose the problem into one more amenable to experimental test? How can I enhance my ability to detect patterns? Where is the boundary between correlation and causality?
- Deduction: How large a role does deduction really play in science? What are some of the more frequent deductive fallacies committed unknowingly by 'logical' scientists?
- Experimental techniques: What seemingly trivial steps can make the difference between an inconclusive experiment and a diagnostic experiment? What troubleshooting procedures have proven effective in all branches of science?
- Objectivity: How much do expectations influence observations? In what ways is objectivity a myth? How can we achieve objective knowledge, in spite of the inescapable subjectivity of individuals?
- Evaluation of evidence: When I think I am weighing evidence rationally, what unconscious values do I employ? How much leverage does prevailing theory exert in the evaluation of new ideas?

- Insight: What are the major obstacles to scientific insight, and how can I avoid them?
- The scientist's world: What issues affect the scientist's interactions with fellow scientists and with society?
- The scientist: What are the *essential* characteristics of successful scientists?

* * *

Thumbnail History of Scientific Methods

What are the essential elements of scientific method, and what are the incidentals? Let's ask history. We can use the Method of Difference (described in Chapter 3): examine changes in the vitality of science as scientific methods evolved. We need to avoid a pitfall: mistaking coincidence for causality (see Chapters 3 and 4).

[Harris, 1970]

To many scientists, the field of history offers little interest. A gap separates the 'two cultures', scientific and literary, and prevents each from appreciating the contributions of the other [Snow, 1964]. Yet even a brief history of the development of scientific methods demonstrates compellingly that:

- communication, particularly access to previous writings, is critical for vitality of science;
- an individual can have a remarkable impact on science -- as an actor or as a mentor;
- we exaggerate our links to the Greeks and to the Italian Renaissance; and
- our 20th century intellectual chauvinism is not justified.

This narrative, like history itself, seems at times to be a string of related, adjacent events rather than an upward evolution toward some objective. Over the past 2500 years, many ingredients of the scientific method ebbed or flowed. More than once, almost all of these elements came together, but they failed to transform because some catalyst was missing.

Fowler [1962] provides a more comprehensive but still concise history of these developments.

* * *

In 399 B.C., a jury of 500 Athenians sentenced Socrates to death. The charges: religious heresy and corrupting the morals of the youth. His crimes: asserting that there is only one God and that people should personally evaluate the meaning of virtue. Perhaps he could have recanted and lived, but the seventy-year-old man chose drinking hemlock over refuting his life's teachings.

His student, Aristocles (Plato), must have been devastated. Plato left Athens and traveled extensively for twelve years. His anguish over the trial ripened into a contempt for democracy and for democratic man:

> "He lives from day to day indulging the appetite of the hour;...His life has neither law nor order; and this distracted existence he terms joy and bliss and freedom; and so he goes on." [Plato, ~427-347 B.C., a]

Finally (and fortunately for the future of Western science) Plato did return to Athens. He taught philosophy just as his mentor had done. One of his students, Euclid, wrote Elements of Geometry, the foundation of geometry for the next twenty-two centuries. Another student, Aristotle, taught Alexander the Great, who fostered the spread of Hellenic science throughout his new empire. The seeds sown by Alexander in Asia flowered throughout Europe more than a thousand years later, catalyzing the 'birth' of the modern scientific method.

Why do I begin this brief history of scientific methods with the death of Socrates and with Plato's response? From Pythagoras to Ptolemy, many individuals built Hellenic science. Yet the heart of this development may be the remarkable mentor-student chain of Socrates-Plato-Aristotle-Alexander. The focal point was not a panorama of historic events, but the response of an individual, Plato, when faced with a choice: should I follow the example of Socrates or should I react against the injustice of society?

Science and the scientific method were not born in Greece. Two criteria for the existence of science -- scientific observation and the collection of facts -- thrived in several pre-Hellenic cultures. Ancient astronomy is the most obvious example: the Mesopotamians in about 3500 B.C., as well as other agricultural cultures at other times, gradually evolved from star-gazing to using the stars and sun for predicting the seasons and eclipses. If technology implies science, should we trace science back to the first use of fire or the first use of tools?

A remarkable number of the key ingredients of scientific methodology were discovered during the Hellenic period:

• Pythagoras, and later Plato, advocated what has become the fundamental axiom of science: the universe is intrinsically ordered and can be understood through the use of reason. Socrates stressed that human intelligence and reason can discover the logical patterns and causal relationships underlying this order. This axiom cannot be proved; we accept it because it is so successful (Killeffer, 1969). Previously, most cultures had interpreted order and law as human concepts that were largely inapplicable to nature.

• Pythagoras identified the relationship between musical notes and mathematics. The Pythagoreans educed that mathematical laws could describe the functioning of nature and the cosmos. Although they did invent geometry, they were unable to develop the mathematical techniques needed to exploit this insight.

• The Hellenic culture, founded on intellectual freedom and love of nature, created a science both contemplative and freer from religious dogma than the preceding and following millennia. The systematic Hellenic investigation of nature, as seen in their geometry, mathematics, astronomy, geography, medicine, and art, may be responsible for our modern Western perception that science had its roots in ancient Greek civilization (Goldstein, 1988). Then, as now, science tested the limits of intellectual freedom. The death of Socrates is proof.

• Aristotle firmly steered Greek science towards rational thought and classification. He honed the blunt tool of deductive logic into the incisive instrument of syllogism. Aristotle also attempted to classify and systematize biological samples that Alexander sent back to him.

• Aristotle also fostered the development of induction, the inference of generalizations from observation: "Now art arises when from many notions gained by experience one universal judgement about a class of objects is produced." [Aristotle, 384-322 B.C.]

Greek science in general, and Aristotle in particular, developed many of the elements of modern scientific method. Yet they neglected verification. Aristotle often succumbed to the rational pitfall of hasty generalization; for example, he claimed that all arguments could be reduced to syllogisms. Greek forays into experimentation and verification, though rare, were sometimes spectacular. In about 240 B.C., for example, Eratosthenes estimated the diameter of the earth, with an error of less than 4%, by measuring the angle of a shadow at Alexandria, when the sun was vertical at Syene. More frequently, however, Greek science ignored experiment and focused instead on the 'higher' skill of contemplative theorizing. Almost two millennia passed before European cultures discarded this bias and thereby embarked on the scientific revolution. Although Aristotle swung the pendulum too far, imparting rigidity to Greek science (Goldstein, 1988), he revealed the potential of deduction and induction.

Science is the Greek word for knowledge. Yet the gift of the Greeks to future science was more a gift of techniques than of facts. Science survived the transition from Greek to Roman culture and the move to Alexandria. But what more can be said of Roman science beyond the observation that its greatest discoveries were the arch, concrete, and improved maps?

* * *

Repeated incursions by nomadic tribes into the boundaries of the Roman Empire eventually overwhelmed the urban Roman civilization. At the same time the appeal of Christian teachings, which provided explanation and solace in the face of increasingly difficult conditions, eventually caused much of the population to embrace the idea that the world of the senses is essentially unreal. Truth lay in the inscrutable plan of God, not in the workings of mathematics. The accompanying eclipse of scientific knowledge and methods went virtually unnoticed. This world-view excluded science, because science requires love of nature and confidence in the world of the senses.

> "The Gothic arms were less fatal to the schools of Athens than the establishment of a new religion, whose ministers superseded the exercise of reason, resolved to treat every question by an article of faith, and condemned the infidel or skeptic to eternal flame." [Gibbon, 1787]

> The scientific nadir came in about 389 A.D.: "In this wide and various prospect of devastation, the spectator may distinguish the ruins of the temple of Serapis, at Alexandria. The valuable library of Alexandria was pillaged, and near twenty years afterwards the appearance of the empty shelves excited the regret and indignation of every spectator whose mind was not totally darkened by religious prejudice. The compositions of ancient genius, so many of which have irretrievably perished, might surely have been excepted from the wreck of idolatry, for the amusement and instruction of succeeding ages." [Gibbon, 1787]

Augustine (354-430 A.D.) was the most eloquent and influential proponent of the new attitude toward science:

> "It is not necessary to probe into the nature of things, as was done by those whom the Greeks call *physici*; nor need we be in alarm lest the Christian should be ignorant of the force and number of the elements - the motion, and order, and eclipses of the heavenly bodies; the form of the heavens; the species and the natures of animals, plants, stones, fountains, rivers, mountains; about chronology and distances; the signs of coming storms; and a thousand other things which those philosophers either have found out or think they have found out...It is enough for the Christian to believe that the only cause of all created things, whether heavenly or earthly, whether visible or in-

> visible, is the goodness of the Creator, the one true God." [St. Augustine, 354-430 A.D., a]

Augustine was probably the major influence on European thought for the next seven centuries. Like other religious mystics before and after him, he turned attention away from rationalism and the senses and toward concern for religion. If the three pillars of wisdom are empiricism, rationalism, and faith (or intuition), then Augustine turned the focus of intellectual thought to the third and previously most neglected of these pillars: intuition, the direct realization of truth by inspiration (Chambliss, 1954). Augustine achieved his insights with the aid of purgation, expecting 'less-disciplined' individuals to accept these insights as dogma. Scientific insights, in contrast, are tested before acceptance. Yet even today scientific insights, once accepted by scientists, are presented to the public as dogma.

In 529 A.D. the Emperor Justinian closed the School of Athens; European science had begun to wane long before. During the long European medieval period of the next six hundred years, technological change virtually ceased. Because technology is an inevitable outgrowth of science, the lack of medieval technological change implies an absence of science.

Augustine had distinguished two types of reason (*ratio*): *sapientia*, the knowledge of eternal things, is the *ratio superior*, while *scientia*, the knowledge of temporal things, is the *ratio inferior* [Fairweather, 1956]. Almost all records from the European medieval period are from the Church, an institution that still followed Augustine's anti-scientific lead. For example, Isidore of Seville's book Etymologies, an early 7th century compilation of knowledge, was influential for 500 years, yet Brehaut [1912] comments on Isidore's 'knowledge':

> "The attitude of Isidore and his time is exactly opposite to ours. To him the supernatural world was the demonstrable one. Its phenomena, or what were supposed to be such, were accepted as valid, while no importance was attached to evidence offered by the senses as to the material."

* * *

Arabs, not Europeans, promoted science throughout the first millennium A.D. Alexander had begun the eastward spread of Greek science. When intellectual freedom waned in the Mediterranean, some scientists and scholars moved to Persia, where it was still encouraged. In the 7th and 8th centuries, the Bedouin tribes of the Arabian Peninsula promulgated Islam throughout the region from Spain to India; they also spread a culture that was remarkably fertile for science.

The Muslim armies were religiously single-minded. They were also tolerant of cultural variations and willing to absorb the heterogeneous cultures that they encountered and conquered. Among the knowledge assimilated were Indian and Babylonian mathematics and the Greek manuscripts. At a time when medieval Europe was turning away from the harshness of worldly affairs, the Muslim were embracing nature's diversity and surpassing the Greeks in applied knowledge. The Arabs adopted Greek scientific methods and knowledge, then added their own observations and came to fresh conclusions. The Arabs were the first to counter the Greek emphasis on contemplation and logic with an insistence on observation.

By the 12th century, Arab science included inexpensive writing paper, medical care (including hospitals), major advances in optics, significant advances in observational astronomy, a highly simplified numeric system, and the equation. The latter two were crucial scientific building blocks. Al-Khwarizmi and other Muslim mathematicians had taken the Babylonian sexagesimal (60-based, e.g. seconds and minutes) and Indian decimal systems and further simplified them into a powerful

mathematical system. This 'Arabic system' included the mathematical use of zero and positional numbers indicating units. Al-Khwarizmi's 'al-jebr' (literally the reducing and recombining of parts), with the simple procedure of changing both sides of the equation by the same amount, allowed complex relationships to be quantified and unknown variables ('x') to be determined in terms of other variables. At last, Pythagoras' dream of a mathematical description of nature was realizable.

These cumulative accomplishments marked the zenith of Arab science. In the 12th century, Muslim science was smothered by the growing consensus that all worthwhile knowledge can be found in the Koran. Science survived through serendipity: after nourishing the flame of science throughout the millennium of anti-science 'Dark Ages' in Europe, the Muslim passed it back to Europe just when a cultural revival there was beginning to crave it.

* * *

The medieval cultural revival of the 12th century began a rediscovery of the most basic scientific foundations. The Catholic Church, sole source of schools and learning, was the epicenter. For example, Peter Abelard used religious reasoning to rediscover the connection between nature and human logic: the universe is logical and ordered because God made it that way; humans were created in God's image so they can decipher the universe's logic. In his book Sic et Non [1122 A.D.], he argued against religious dogmatism and for personal critical evaluation:

> "All writings belonging to this class [of scriptural analysis] are to be read with full freedom to criticize, and with no obligation to accept unquestioningly . . . These questions ought to serve to excite tender readers to a zealous inquiry into truth and so sharpen their wits. The master key of knowledge is, indeed, a persistent and frequent questioning. . . By doubting we come to examine, and by examining we reach the truth."

The scientific renaissance began in the 12th-century cathedral schools, particularly the School of Chartres [Goldstein, 1988]. By the early 13th century, the surge of knowledge had moved to the first universities, such as those in Paris, Oxford, and Salerno. Yet, in the brief period surrounding the construction of the cathedral of Chartres, its school made several impressive innovations:

- establishment of the natural sciences as areas of study at least as important as liberal arts;
- creation of the first substantial library of science since Roman times, with a particular emphasis on collecting ancient scientific writings;
- reintroduction of the Pythagorean idea of a mathematically ordered structure of the universe; and
- search for causality throughout nature, based on the idea that "nature is intelligible for the human mind precisely because both proceed according to the same inherent rational law" [Goldstein, 1988].

The architects of the new science at the School of Chartres were Thierry of Chartres and his student William of Conches. Thierry laid the groundwork by establishing religious justifications for the study of nature. He asked, "Given God, how do we prove it?" and he encouraged scientific contribution to this goal. William of Conches [~1150 A.D.] was less cautious:

> "To seek the 'reason' of things and the laws governing their production is the great task of the believer and one which we should discharge together, bound by our curiosities into a fraternal enterprise."

Inevitably, this fundamental modification of perception aroused the fear and anger of conservatives. Inevitably, conservatives attempted to use the Church to prevent the change, by arguing that this altered perception violated fundamental principles of the Church. The battle that began then -- as a conflict between two religious views of nature -- continues even today, couched as a conflict between science and religion.

> "Science and religion, religion and science, put it as I may they are two sides of the same glass, through which we see darkly until these two, focusing together, reveal the truth." [Buck, 1962]

The enemy of science then and today is not religion, any more than the enemy of science during Plato's day was democracy. Both the Christian religion and democratic laws had seemed threatening when they were introduced. Later, each became the weapon wielded by conservatives to protect themselves from the fear engendered by scientific change. Unlike the conservatives and religious zealots, scientists greet claims of 'absolute truth' with skepticism. Revelation is eventually seen as naïveté, for all understandings evolve and improve.

The *status quo* will always be used to challenge scientific change.

* * *

At about the same time that the School of Chartres was rediscovering Greek knowledge with their own pitifully small library, Europeans encountered the entirety of Greek and Arab scientific knowledge on several fronts. In Spain the long civil war between Christians and Muslims led to capture of Muslim cities, and the Christian king Alfonso VII established a center in Toledo for the study of Islamic science. The Crusaders also found libraries rich in Greek manuscripts, particularly during the capture of Constantinople in 1204. When the emerging European spirit of scientific enquiry encountered potential answers in the form of Greek and Arab scientific writings, translators were kept busy for more than a century.

Eight hundred years later as I write this, war between Western Christians and Arab Muslims has flared again, and the Arab gift to the west of practical applied science is returning to Iraq in the form of high-technology weapons.

Much of the Arab science was not fully absorbed by the Europeans for centuries. Scientific knowledge was only a part of the Islamic gift to the Europeans. The Islamic pleasure and curiosity in observing nature's diversity spearheaded a 12th-century cultural and scientific renaissance of intellectual and sensual liberation [Goldstein, 1988]. This renaissance was exemplified by Robert Grossteste (1175-1253), once chancellor of Oxford, and his student Roger Bacon (1214-1294 A.D.). Grossteste gave the first relatively complete description of modern scientific method, including induction, experimentation, and mathematics [Crombie, 1953]. Bacon argued that it is necessary to combine mathematical analysis with empirical observation and that experiments should be controlled. More than two centuries before the technological insights of Leonardo da Vinci, Roger Bacon [~1270 A.D.] foresaw the potential technological results of scientific method:

> "Great ships and sea-going vessels shall be made which can be guided by one man and will move with greater swiftness than if they were full of oarsmen. It is possible that a car shall be made which will move with inestimable speed, and the motion will be without the help of any living creature. . . A device for flying shall be made such that a man sitting in the middle of it and turning a crank shall cause artificial wings to beat the air after the manner of a flying bird. Similarly, it is possible to construct a small-sized instrument for elevating and depressing great weights . . . It is possible also that devices can be made whereby, without bodily danger, a man may walk on the bottom of the sea or of a river."

Grosseteste and Bacon were prophets, not flag-bearers, of the coming new science. Their emphasis on observational, empirical science was overshadowed by a prevailing respect for authority that fostered acceptance of the ancient writings [Haskins, 1927]. The scholastic Albertus Magnus [~1250 A.D.] responded with the still-familiar rebuttal: "experience is the best teacher in all such things." Their contemporary Thomas Aquinas was more persuasive; he created a mainstream scholastic attitude that empiricism and rationalism should have the more limited scope of serving religion.

The scholastic approach of combining reason and faith was more scientifically effective than the Islamic approach of accepting diverse perspectives without requiring intellectual consistency among them. By the beginning of the 14th century, the young European science had already surpassed its Greek and Arab parents, partly because earlier Christian theological arguments had fostered a rationalist, logical style of evaluating abstract concepts. Yet a strong tradition of mysticism was able to exist side-by-side with the rationalist school of the Scholastics. The mystic tradition was less scientifically efficient than more rational science, because it encompassed research on invisible powers. Yet the centuries of alchemical research encouraged creativity and patient observation and eventually gave birth to modern chemistry.

In the 15th and 16th centuries, an Italian Renaissance gained momentum and the pace of change increased. Begun as a revival of interest in Greek and Roman literature, it rejected the otherworldly traditions of the previous millennium and embraced the love of nature and study of nature, at first through art and later also through science. Leonardo da Vinci (1452-1519) exemplifies the intimate relationship of art to science in this period, as well as the age's spirit of curiosity. The synergy of curiosity about nature, medieval rationalism, and empiricism led to an age of exploration and to the scientific revolution.

[Harris, 1970]

> "The scientific revolution began in curiosity, gained momentum through free inquiry, [and] produced its first fruits in knowledge of the material universe." [Chambliss, 1954]

> "The condition most favorable to the growth of science in the sixteenth and seventeenth centuries was the increasing number of men who were drawn into intellectual pursuits. Genius is like a fire; a single burning log will smolder or go out; a heap of logs piled loosely together will flame fiercely. . . But the establishment of strong governments, insuring at least domestic peace, the accumulation of wealth followed by the growth of a leisure class, the development of a secular, sanguine culture more eager to improve this world than anxious about the next, and above all, the invention of printing, making easier the storing, communication, and dissemination of knowledge, led naturally to the cultivation and hence to the advancement of science." [Smith, 1930]

There were scientific setbacks in these centuries, but the acceleration of science could not be stopped. In 1543, European science took a quantum leap forward into the scientific revolution, as the result of publication of three remarkable books:

• Archimedes' book on mathematics and physics was translated from the Greek and became widely read for the first time;

• The Structure of the Human Body, a book of Andreas Vesalius' anatomical drawings, provided the first accurate look at human anatomy;

• The Revolution of the Heavenly Spheres, by Nicolaus Copernicus, presented the concept of a heliocentric cosmology and set the scientific revolution in motion, as its author lay on his deathbed.

Giordano Bruno (1473-1543) advocated this Copernican universe and was burned at the stake. A century later Galileo strongly argued for a Copernican universe. He was tried by the church and threatened with excommunication, he was forced to recant, and he spent the rest of his life under house arrest. Later scientists, particularly Kepler and Newton, concluded the battle with less adverse personal impact. Bronowski [1973] calls Galileo the "creator of the modern scientific method" because in 1609-1610 he designed and built a 30-power telescope, used it for astronomical observations, and published the result. I see Galileo not as the creator but as one who exemplifies an important phase in the evolution of modern scientific method.

Galileo valued experimental verification of ideas. In the 17th century, Francis Bacon, René Descartes, and others succeeded in steering science away from mysticism and confining scientific research to topics that are verifiable, by either the senses or deduction. Indeed, even the 17th-century scientific genius Isaac Newton devoted part of his life to alchemy. When researchers adopted the pragmatic attitude of giving priority to what is observable with the senses, they took one of the final steps in development of modern scientific method.

* * *

The early 17th century saw a watershed collision of two philosophies of scientific method: deduction and experimentation. René Descartes' [1637] book Discourse on Method emphasized mathematical deduction and laid out the following four principles of his scientific method:

• "never accept anything as true if I had not evident knowledge of its being so. . .

• divide each problem I examined into as many parts as was feasible. . .

• direct my thoughts in an orderly way; beginning with the simplest objects. . .

• make throughout such complete enumerations that I might be sure of leaving nothing out."

In contrast, Francis Bacon's [1620] book Novum Organum sought to establish a new empirical type of science. He argued compellingly that science cannot be confined to either deduction or observation; one must use a combination of experiment and hypothesis, testing hypotheses empirically.

> "All true and fruitful natural philosophy hath a double scale or ladder, ascendent and descendent, ascending from experiments to the invention of causes, and descending from causes to the invention of new experiments." [Bacon, 1561-1626]

Both approaches had strengths and weaknesses, and both contributed to modern scientific method. Bacon, who was not a working scientist, failed to realize the importance of intuition in creating hypotheses and of judgment in rejecting most hypotheses so that only a subset need be tested. Descartes sought to confine science to those areas in which mathematics could yield 'certainty':

> "Science in its entirety is true and evident cognition. He is no more learned who has doubts on many matters than the man who has never thought of them; nay he appears to be less learned if he has formed wrong opinions on any particulars. Hence it

> were better not to study at all than to occupy one's self with objects of such difficulty, that, owing to our inability to distinguish true from false, we are forced to regard the doubtful as certain; for in those matters any hope of augmenting our knowledge is exceeded by the risk of diminishing it. Thus . . . we reject all such merely probable knowledge and make it a rule to trust only what is completely known and incapable of being doubted." [Descartes, ~1629]

This deductive dogmatism is incompatible with almost all of modern science; even theoretical deductive physics begins with unproven premises. In the 17th century, however, the outcome of the battle over the future direction of science could not be predicted.

Antoine Arnauld [1662], in an influential book on logic, presented a pragmatic approach to scientific and other judgment: rational action, like gambling, is based not on Cartesian certainty but on consideration of the probabilities of the potential outcomes. Isaac Newton [1687] reinforced the Cartesian perspective on science with his book Principia Mathematica. Considered by some to be the most important scientific book in history, Principia established a new paradigm of the physics of motion, drawing together a very wide suite of observations into a rigorous mathematical system. Newton was primarily a theoretician, not an empiricist, but he eagerly used data collected by others.

Eventually, the conflict was resolved: with the edges of the road defined, a middle way could be trod. John Locke argued persuasively that experimental science is at least as important as Cartesian deduction. Locke became known as the 'father of British empiricism.' 'Champion of empiricism' is probably a more appropriate epithet, however, for Locke made no important scientific discoveries. Locke provided the needed scientific compromise: certainty is possible in mathematics, but most scientific judgments are based on probable knowledge rooted in controlled experiments. Each person must evaluate open-mindedly the evidence and make a personal judgment.

> "The mechanical world view is a testimonial to three men: Francis Bacon, René Descartes, and Isaac Newton. After 300 years we are still living off their ideas." [Rifkin, 1980]

> Isaac Newton [1676] wrote to Robert Hooke, "If I have seen a little further it is by standing on the shoulders of Giants."

> Bernard of Chartres [~1150] wrote, "We are like dwarfs sitting on the shoulders of giants; we see more things, and things that are further off, than they did -- not because our sight is better, or because we are taller than they were, but because they raise us up and add to our stature by their gigantic height."

* * *

Remarkably, scientific method has changed very little in the last three centuries.

Archimedes [~287-212 B.C.], emphasizing the power of the lever, boasted, "Give me a place to stand on and I can move the earth." Of course, no lever is that strong. Even the 300-ton blocks of Egyptian and Middle American pyramids were beyond the strength of individual, rigid levers; recent research suggests the possibility that many flexible bamboo levers could have shared and distributed each load [Cunningham, 1988].

Three hundred years ago, the suite of scientific levers was completed. The world began to move in response.

* * *

Myth of a Scientific Method

> "The unity of science, which is sometimes lost to view through immersion in specialist problems, is essentially a unity of method." [Russell, 1938]

> "But on one point I believe almost all modern historians of the natural sciences would agree. . .There is no such thing as *the* scientific method." [Conant, 1947]

Our brief examination of the history of science suggests that trial and error have refined the following elements of modern, successful scientific method:

> Facts are collected by carefully controlled experiments. Based on these facts, verifiable hypotheses are proposed, objectively tested by further experiments, and thereby proven or discarded.

This view of scientific method was universally embraced in the 19th century, and it is still popular. Most scientists would probably comment that this two-sentence description is necessarily a bit simplistic but is about right. They would replace the word 'facts', however, by 'observations', because they recognize that science is too dynamic for any data or ideas to be considered as immortal facts.

Philosophers of science universally reject this view of scientific method. They emphasize that objectivity is a myth, that experimental observations are inseparable from theories, and that hypothesis tests seldom cause rejection of a hypothesis and cannot prove a hypothesis. Furthermore, it is impossible to define a single scientific method shared by all scientists; the sciences and scientists are far too heterogeneous. Most philosophers of science conclude that the term 'scientific method' should be abandoned.

> "Scientists are people of very dissimilar temperaments doing different things in very different ways. Among scientists are collectors, classifiers, and compulsive tidiers-up; many are detectives by temperament and many are explorers; some are artists and others artisans." [Medawer, 1967]

Both scientists and philosophers seek universal concepts, but scientists often settle for less: an idea may still be considered useful even if it does not fit all relevant data. We scientists can abandon the idea of '*the* scientific method' but still embrace the concept of 'scientific methods' -- a suite of logical techniques, experimental techniques, principles, evaluation standards, and even ethical standards. Unlike Francis Bacon and Renè Descartes, modern scientists can select from this suite without rejecting the constructs of those who choose to use different methods. We must, however, know the limitations of both our methods and theirs.

* * *

Scientific Methods

Two of the most fundamental tools in the scientific toolbox are data and concepts. So basic is the distinction between the two, that nearly all publications confine data and interpretations to separate sections. Clearly, interpretations depend on data; less obviously, all data collection involves concept-based assumptions (Chapter 6).

Explanatory concepts can be given different labels, depending on our confidence in their reliability. A **law** is an explanation in which we have the greatest confidence, based on a long track record of confirmations. A **theory**, for most scientists, denotes an explanation that has been confirmed sufficiently to be generally accepted, but which is less firmly established than a law. An **axiom** is a concept that is accepted without proof, perhaps because it is obvious or universally accepted (e.g., time, causality) or perhaps to investigate its implications. A **hypothesis** is an idea that is still in the process of active testing; it may or may not be correct. **Models** are mathematical or conceptual hypotheses that provide useful perspectives in spite of recognized oversimplification. Whereas laws and theories are relatively static, hypothesis formation, testing, and evaluation are the dynamic life of science.

Laws, theories, and hypotheses also differ in generality and scope. Theories tend to be broadest in scope (e.g., the theories of relativity and of natural selection); most provide a unified perspective or logical framework for a variety of more specific and more limited laws and hypotheses. All three are generalizations; rarely do they claim to predict the behavior of every particular case, because they cannot encompass all variables that could be relevant. Most laws are expected to be universal in their applicability to a specified subset of variables, but some are permitted exceptions. For example, the geological law of original horizontality states that *nearly* all sediments are initially deposited *almost* horizontally. Hypotheses have not fully bridged the gap between the particular and the universal; most are allied closely with the observations that they attempt to explain.

Researchers do not take these terms too seriously, however. The boundaries between these three categories are flexible, and the terms may be used interchangeably.

Hypotheses, theories, and laws are explanations of nature, and explanations can be qualitative or quantitative, descriptive or causal (Chapter 3). Most explanations involve variables -- characteristics that exhibit detectable and quantifiable changes (Chapter 2) -- and many explanations attempt to identify relationships among variables (Chapter 3).

All scientific concepts must be testable -- capable of confirmation or refutation by systematic reality checking. Uncertainty is inherent not only to explanatory concepts, but also in the terms describing concept testing: confirmation, verification, validity, reliability, and significance. Scientific confirmation does not establish that an idea must be correct, or even that it is probably correct. Confirmation is merely the demonstration that a hypothesis is consistent with observations, thereby increasing confidence that the hypothesis is correct.

Some concepts can be tested directly against other, more established concepts by simple logical deduction (Chapter 4). More often, we need to investigate the hypothesis more indirectly, by identifying and empirically testing predictions made by the concept. **Data** are the experimental observations, or measurements, that provide these tests.

The interplay between hypothesis and data, between speculation and reality checking, is the heart of scientific method. Philosophers of science have devoted much analysis to the question of how hypothesis and data interact to create scientific progress. In the latter half of the twentieth century, the leading conceptualization of the scientific process has been the hypothetico-deductive method. Popper [1959, 1963], Medawer [1969], Achinstein [1985], and many others have provided perspectives on what this method is and what it should be. Most suggest that the gist of this **method of hypothesis** is the following:

> Scientific progress is achieved by interplay between imagination and criticism. Hypotheses are the key, and hypothesis formation is a creative act, not a compelling product of observations. After developing a hypothesis, the scientist temporarily as-

> sumes that it is correct, then determines its logical consequences. This inference may be deductive, a necessary consequence of the hypothesis, or inductive, a probable implication of the hypothesis. To be fruitful, this inference must generate a testable prediction of the hypothesis. An experiment is then undertaken to confirm or refute that prediction. The outcome affects whether the hypothesis is retained, modified, or refuted.

This method of hypothesis is the crux of scientific method, but scientific progress need not be quite as linear as shown. Hypotheses can be generated at any stage. Most die virtually immediately, because they are incompatible with some well-established observations or hypotheses. A single hypothesis may yield multiple predictions: some useless, many testable by a brief search of published experiments, some requiring an experiment that is infeasible, and few leading to actual experiments.

The insistence on verifiability, or its converse -- falsifiability, limits the scope of science to the pursuit of verifiably **reliable knowledge**. Reliability is, however, subjective (see Chapters 6 and 7), and hypothesis tests are seldom as conclusive as we wish. Though a hypothesis test cannot prove a hypothesis, some scientists (especially physicists) and many philosophers claim that it can at least disprove one. This argument, however, holds only for deductive predictions. More often, the test is not completely diagnostic, because assumptions buffer the hypothesis from refutation. Many hypotheses are abandoned without being refuted. Others are accepted as reliable without proof, if they have survived many tests; we cannot work effectively if we constantly doubt everything.

* * *

Is there a scientific method? The answer depends on whether one is a lumper or a splitter. Certainly the method of hypothesis is central to nearly all science, but scientific techniques and style depend both on the problem investigated and on individual taste.

For some, like Francis Bacon or Thomas Edison, experimentation is exploration; interpretations will inevitably follow. Trial and error, with *many* trials, is the method used by Edison, the medieval alchemists, and modern seekers of high-temperature superconductors. Others, like Aristotle, employ the opposite approach: develop an idea, then experimentally demonstrate its validity. A few, like René Descartes or Immanuel Kant, deduce the implications of premises. Many more, like Galileo, make predictions based on a hypothesis and empirically test those predictions. For most, each of these approaches is sometimes useful.

Research style is also fluid. At one extreme is Leonardo da Vinci, fascinated by everything he saw. Mohammad Ali, in describing himself, also described this research style: “Dance like a butterfly; sting like a bee.” At the other extreme is the Great Pyramid style -- systematically and possibly laboriously undertake multiple experiments in the same field, until the final foundation is unshakeable. Charles Darwin used this strategy for establishing his theory of evolution, except that he compiled evidence rather than experiments.

The scientific method is both very liberal and very conservative: any hypothesis is worthy of consideration, but only those that survive rigorous testing are incorporated into the framework of reliable knowledge. The scientific method is incredibly versatile, both in the range of knowledge amenable to its investigation and in the variety of personal scientific styles that it fosters and embraces. Invariably, however, it demands an intriguing and challenging combination: creativity plus skepticism.

Chapter 2: Variables

A variable is a characteristic that exhibits detectable changes, either regionally or temporally. Implicit in this concept of change is influence by something else: Newtonian dynamics show us that movement in itself does not imply an external force -- change in movement does. Thus scientists are seldom concerned with a single variable; more often we seek patterns among variables. This chapter focuses on one variable at a time, thereby setting the stage for considerations in the next chapter of relationships among variables.

Variables, measurements, and quantification are related components of the foundation of science. Characterizing a variable requires measurements, and measurements require prior quantification. Each variable has an associated measurement type:

- **Nominal** measurements classify information and count the frequency of observations within each class. An example is tabulation of the numbers of protons and neutrons in an atom.

- **Ordinal** measurements specify order, or relative position. Ordinal scales may be objectively determined (e.g., the *k*, *l*, and *m* electron shells), subjectively determined but familiar (e.g., youth, young adult, adult, old age), or subjectively invented for a particular experiment (e.g., social science often uses scales such as this: strongly agree (+2), agree (+1), not sure (0), disagree (-1), strongly disagree (-2)).

- **Interval** and **ratio** scales permit measurements of distance along a continuum, with determinable distance between data. The scales involve either real (fractional) or integer (counting) numbers. They differ in that only ratio scales have a true, or meaningful, zero, permitting determination of ratios between data measurements. For example, temperatures are measured with an interval scale, whereas length is a ratio scale. To refer to one temperature as twice that of another is pointless, whereas it is valid to say that one object is twice as long as another.

The initial quantification of any variable is challenging, for we seek a scale that is both measurable and reliable. Soon, however, that quantification is taken for granted. Finding a way to quantify some variable tends to be more of a problem in the social sciences than in the physical sciences.

* * *

Statistics

Statistics are pattern recognition transformed, from a qualitative guess about what may be, into a quantitative statement about what probably is.

A decade ago, scientific statistics usually required either complex number crunching or simplifying approximations. Since then, computers have revolutionized our approach to statistics. Now standard statistical techniques are available on most personal computers by simply choosing an option, and programs for even the most sophisticated statistical techniques are available from books such as Numerical Recipes [Press et al., 1988]. With easy access comes widespread misuse; one can use various statistical routines without learning their assumptions and limitations.

Statistics help us both to solve single-variable problems (this chapter) and to accomplish multivariate pattern recognition (next chapter). Neither chapter is a substitute for a statistics course or statistics book; no proofs or derivations are given, and many subjects are skipped. Statistics books

present the forward problem of looking at the statistical implications of each of many groups of initial conditions. The scientist more often faces the inverse problem: the data are in-hand, and the researcher wonders which of the hundreds of techniques in the statistics book is relevant.

Efficiency is a key to productive science. Statistics and quantitative pattern recognition increase that efficiency in many ways: optimizing the number of measurements, extracting more information from fewer observations, detecting subtle patterns, and strengthening experimental design. Statistics may even guide the decision on whether to start a proposed experiment, by indicating its chance of success. Thus, it is a disservice to science to adopt the attitude of Rutherford [Bailey, 1967]: "If your experiment needs statistics, you ought to have done a better experiment."

These two chapters introduce some of the statistical methods used most frequently by scientists, and they describe key limitations of these techniques. Rather than an abridged statistics book, the chapters are more an appetizer, a ready reference, an attempt at demystifying a subject that is an essential part of the scientific toolbox.

* * *

Errors

All branches of science use numerical experimental data, and virtually all measurements and all experimental data have **errors** -- differences between measurements and the *true* value. The only exceptions that I can think of are rare integer data (e.g., how many of the subjects were male?); real numbers are nearly always approximate. If a measurement is repeated several times, measurement errors are evident as a measurement scatter. These errors can hide the effect that we are trying to investigate.

* * *

Errors do not imply that a scientist has made mistakes. Although almost every researcher occasionally makes a mathematical mistake or a recording error, such errors are sufficiently preventable and detectable that they should be extremely rare in the final published work of careful scientists. Checking one's work is the primary method for detecting personal mistakes. Scientists vary considerably in how careful they are to catch and correct their own mistakes.

How cautious should a scientist be to avoid errors? A scientist's rule of thumb is that *interpretations can be wrong, assumptions can be wrong, but there must be no data errors due to mistakes. The care warranted to avoid errors is proportional to the consequences of mistakes*. A speculation, if labeled as such, can be wrong yet fruitful, whereas the incorrect announcement of a nonprescription cure for cancer is tantamount to murder.

> Physicist George F. Smoot set a standard of scientific caution: for 1 1/2 years he delayed announcement of his discovery of heterogeneities in the background radiation of the universe, while his group searched avidly for any error in the results. He knew that the discovery provided critical confirmation of the Big Bang theory, but he also knew that other scientists had mistakenly claimed the same result at least twice before. Consequently, he made the following standing offer to members of his research team: to anyone who could find an error in the data or data analysis, he would give an air ticket to anywhere in the world. [5/5/92 New York Times]
>
> I imagine that he was also careful to emphasize that he was offering a round-trip ticket, not a one-way ticket.

Both scientists and non-scientists have recognized the constructive role that error can play:

"Truth comes out of error more readily than out of confusion." [Francis Bacon, 1620]

"The man who makes no mistakes does not usually make anything." [Phelps, 1899]

Incorrect but intriguing hypotheses can be valuable, because the investigations that they inspire may lead to a discovery or at least show the way toward a better hypothesis (Chapter 7). Humphrey Davy [1840] said, "The most important of my discoveries have been suggested by my failures." More rarely, incorrect evidence can inspire a fruitful hypothesis: Eldredge and Gould's [1972] seminal reinterpretation of Darwinian evolution as punctuated equilibrium initially was based on inappropriate examples [Brown, 1987].

Errors are most likely to be detected upon first exposure. Once overlooked, they become almost invisible. For example, if an erroneous hypothesis passes initial tests and is accepted, it becomes remarkably immune to overthrow.

"One definition of an expert is a person who knows all the possible mistakes and how to avoid them. But when we say that people are 'wise' it's not usually because they've made every kind of mistake there is to make (and learned from them), but because they have stored up a lot of simulated scenarios, because their accumulated quality judgments (whether acted upon or not) have made them particularly effective in appraising a novel scenario and advising on a course of action." [Calvin, 1986]

* * *

Errors arise unavoidably: unrecognized variations in experimental conditions generate both so-called 'random' errors and systematic errors. The researcher needs to know the possible effects of errors on experimental data, in order to judge whether or not to place any confidence in resulting conclusions. Without knowledge of the errors, one cannot compare an experimental result to a theoretical prediction, compare two experimental results, or evaluate whether an apparent correlation is real. In short, the data are nearly useless.

* * *

Precision > Accuracy > Reliability

The terms precision, accuracy, reliability, confidence, and replicatability are used interchangeably by most non-scientists and are even listed by many dictionaries as largely synonymous. In their scientific usage, however, these terms have specific and important distinctions.

Errors affect the precision and accuracy of measurements. **Precision** is a measure of the scatter, dispersion, or **replicatability** of the measurements. Low-precision, or high-scatter, measurements are sometimes referred to as noisy data. Smaller average difference between repeat (replicate) measurements means higher precision. For example, if we measure a sheet of paper several times with a ruler, we might get measurements such as 10.9", 11.0", 10.9", and 11.1". If we used a micrometer instead, we might get measurements such as 10.97", 10.96", 10.98", and 10.97". Our estimates show random variation regardless of the measuring device, but the micrometer gives a more precise measurement than does the ruler. If our ruler or micrometer is poorly made, it may yield measurements that are consistently offset, or *systematically biased*, from the true lengths. **Accuracy** is the extent to which the measurements are a reliable estimate of the 'true' value. Both random errors and systematic biases reduce accuracy.

Reliability is a more subjective term, referring usually to interpretations but sometimes to measurements. Reliability is affected by both precision and accuracy, but it also depends on the validity of any assumptions that we have made in our measurements and calculations. Dubious assumptions, regardless of measurement precision and accuracy, make interpretations unreliable.

* * *

Random and Systematic Errors

Random errors are produced by multiple uncontrolled and usually unknown variables, each of which has some influence on the measurement results. If these errors are both negative and positive perturbations from the *true* value, and if they have an average of zero, then they are said to affect the precision of replicate measurements but they do not bias the average measurement value.

If the errors average to a nonzero value, then they are called systematic errors. A constant systematic error affects the accuracy but not the precision of measurements; a variable systematic error affects both accuracy and precision. Systematic errors cause a shift of individual measurements, and thus also of the average measured value, away from the true value. Equipment calibration errors are a frequent source of systematic errors. Inaccurate calibration can cause all values to be too high (or low) by a similar percentage, a similar offset, or both. An example of a systematic percentage bias is plastic rulers, which commonly are stretched or compressed by about 1%. An example of an offset bias is using a balance without zeroing it first. Occasionally, systematic errors may be more complicated. For example, a portable alarm clock may be set at a slightly incorrect time, run too fast at first, and run too slowly when it is about ready for rewinding.

Both random and systematic errors are ubiquitous. In general, 'random errors' only appear to be random because we have no ability to predict them. If either random or systematic errors can be linked to a causal variable, however, it is often possible to remove their adverse effects on both precision and accuracy.

> One person's signal is another person's noise, I realized when I was analyzing data from the Magsat satellite. Magsat had continuously measured the earth's magnetic field while orbiting the earth. I was studying magnetism of the earth's crust, so I had to average out atmospheric magnetic effects within the Magsat data. In contrast, other investigators were interested primarily in these atmospheric effects and were busily averaging out crustal 'contamination.'

Random errors can be averaged by making many replicate, or repeat, measurements. **Replicate measurements** allow one to estimate and minimize the influence of random errors. Increasing the number of replicate measurements allows us to predict the true value with greater confidence, decreasing the **confidence limits** or range of values within which the true value lies. Increasing the number of measurements does not rectify the problem of systematic errors, however; experimental design must anticipate such errors and attenuate them.

* * *

Representative Sampling

Most experiments tackle two scientific issues -- reducing errors and extrapolating from a sample to an entire population -- with the same technique: representative sampling. A **representative sample** is a small subset of the overall population, exhibiting the same characteristics as that population. It is also a prerequisite to valid statistical induction, or quantitative generalization. Nonrepresentative sampling is a frequent pitfall that is usually avoidable. Often, we seek patterns applicable to a broad population of events, yet we must base this pattern recognition on a small subset of the

population. If our subset is representative of the overall population, if it exhibits similar characteristics to any randomly chosen subset of the population, then our generalization *may* have applicability to behavior of the unsampled remainder of the population. If not, then we have merely succeeded in describing our subset.

Representative sampling is essential for successful averaging of random errors and avoidance of systematic errors, or bias. *Random sampling achieves representative sampling*. No other method is as consistently successful and free of bias. Sometimes, however, random sampling is not feasible. With random sampling, every specimen of the population should have an equal chance of being included in the sample. Every specimen needs to be numbered, and the sample specimens are selected with a random number generator. If we lack access to some members of the population, we need to employ countermeasures to prevent biased sampling and consequent loss of generality. Stratification is such a countermeasure.

Stratification does not attempt random sampling of an entire population. Instead, one carefully selects a subset of the population in which a primary variable *is* present at a representative level. Stratification is only useful for assuring representative sampling if the number of primary variables is small. Sociologists, for example, cannot expect to find and poll an 'average American family'. They can, however, investigate urban versus rural responses while confining their sampling to a few geographical regions, if those regions give a stratified, representative sample of both urban and rural populations.

For small samples, stratification is actually more effective in dealing with a primary variable than is randomization: stratification deliberately assures a representative sampling of that variable, whereas randomization only approximately achieves a representative sample. For large samples and many variables, however, randomization is safer. Social sciences often use a combination of the two: stratification of a primary variable and randomization of other possible variables [Hoover, 1988]. For example, the Gallup and Harris polls use random sampling within a few representative areas.

> In 1936, the first Gallup poll provided a stunning demonstration of the superiority of a representative sample over a large but biased sample. Based on polling twenty million people, the Literary Digest projected that Landon would defeat Roosevelt in the presidential election. The Literary Digest poll was based on driver's license and telephone lists; only the richer segment of the depression-era population had cars or telephones. In contrast, George Gallup predicted victory for Roosevelt based on a representative sample of only ten thousand people.

The concept of random sampling is counterintuitive to many new scientists and to the public. A carefully chosen sample seems preferable to one selected randomly, because we can avoid anomalous, rare, and unusual specimens and pick ones exhibiting the most typical, broad-scale characteristics. Unfortunately, the properties of such a sample probably cannot be extrapolated to the entire population. Statistical treatment of such data is invalid. Furthermore, sampling may be subconsciously biased, tending to yield results that fulfill the researcher's expectations and miss unforeseen relationships (Chapter 6). Selective sampling *may* be a valid alternative to random sampling, if one confines interpretations to that portion of the population for which the sample is a representative subset.

Even representative sampling cannot assure that the results are identical to the behavior of the entire population. For example, a single coin flip, whether done by hand or by a cleverly designed unbiased machine, will yield a head or a tail, not 50% heads and 50% tails. The power of random sampling is that it can be analyzed reliably with quantitative statistical techniques such as those described in this chapter, allowing valid inferences about the entire population. Often these inferences

are of the form '*A* probably is related to *B*, because within my sample of *N* specimens I observe that the A_i are correlated with B_i.'

* * *

Replication and Confirmation

The terms **replicatability** and **reproducibility** are often used to refer to the similarity of replicate measurements; in this sense they are dependent only on the precision of the measurements. Sometimes *replicatability* is used in the same sense as *replication*, describing the ability to repeat an entire experiment and obtain substantially the same results. An experiment can fail to replicate because of a technical error in one of the experiments. More often, an unknown variable has different values in the two experiments, affecting them differently. In either case the failure to replicate transforms conclusions. Identifying and characterizing a previously unrecognized variable may even eclipse the original purpose of the experiments.

Replication does not imply duplication of the original experiment's precision and accuracy. Indeed, usually the second experiment diverges from the original in design, in an attempt to achieve higher precision, greater accuracy, or better isolation of variables. Some [e.g., Wilson, 1952] say that one *should not* replicate an experiment under exactly the same conditions, because such experiments have minor incremental information value. Exact replication also is less exciting and less fundamental than novel experiments.

If the substantive results (not the exact data values but their implications) or conclusions of the second experiment are the same as in the first experiment, then they are **confirmed**. Confirmation does not mean proved; it means strengthened. Successful replication of an experiment is a confirmation. Much stronger confirmation is provided by an experiment that makes different assumptions and different kinds of measurements than the first experiment, yet leads to similar interpretations and conclusions.

In summary, precision is higher than accuracy, because accuracy is affected by both precision and systematic biases. Accuracy is higher than reliability, because reliability is affected not just by measurement accuracy but also by the validity of assumptions, simplifications, and possibly generalizations. Reliability is increased if other experiments confirm the results.

* * *

Probability

Probability is a concern throughout science, particularly in most social sciences, quantum physics, genetics, and analysis of experiments. Probability has a more specific meaning for mathematicians and scientists than for other people. Given a large number of experiments, or trials, with different possible outcomes, probability is the proportion of trials that will have one type of outcome. Thus the sum of probabilities of all possible outcomes is one.

Greater probability means less uncertainty, and one objective of science is to decrease uncertainty, through successful prediction and the recognition of orderly patterns. Induction (Chapter 3), which is a foundation of science, is entirely focused on determining what is *probably* true. Only by considering probability can we evaluate whether a result could have occurred by chance, and how much confidence to place in that result.

"Looking backwards, any particular outcome is always highly improbable" [Calvin, 1986]. For example, that I am alive implies an incredibly improbable winning streak of birth then reproduction

before death that is several hundred million years long. Yet I do not conclude that I probably am not alive. The actual result of each trial will be either occurrence or nonoccurrence of a specific outcome, but our interest is in proportions for a large number of trials.

The most important theorem of probability is this: when dealing with several independent events, the probability of all of them happening is the product of the individual probabilities. For example, the probability of flipping a coin twice and getting heads both times is $1/2 \bullet 1/2 = 1/4$; the chance of flipping a coin and a die and getting a head plus a two is $1/2 \bullet 1/6 = 1/12$. If one has already flipped a coin twice and gotten two heads, the probability of getting heads on a third trial and thus making the winning streak three heads in a row is $1/2$, not $1/2 \bullet 1/2 \bullet 1/2 = 1/8$. The third trial is independent of previous results.

Though simple, this theorem of multiplicative probabilities is easy to misuse. For example, if the probability of getting a speeding ticket while driving to and from work is 0.05 (i.e., 5%) per round trip, what is the probability of getting a speeding ticket sometime during an entire week of commuting? The answer is not $.05 \bullet .05 \bullet .05 \bullet .05 \bullet .05 = .0000003$; that is the probability of getting a speeding ticket on every one of the five days. If the question is expressed as "what is the probability of getting *at least one* speeding ticket", then the answer is $1-0.95^5=0.226$, or 1 minus the probability of getting no speeding tickets at all.

Often the events are not completely independent; the odds of one trial are affected by previous trials. For example, the chance of surviving one trial of Russian roulette with a 6-shot revolver is $5/6$; the chance of surviving two straight trials (with no randomizing spin of the cylinders between trials) is $5/6 \bullet 4/5 = 2/3$.

Independence is the key to assuring that an undesirable outcome is avoided, whether in a scientific research project or in everyday life. The chance of two independent rare events occurring simultaneously is exceedingly low. For example, before a train can crash into a station, the engineer must fail to stop the train (e.g., fall asleep) *and* the automatic block system must fail. If the chance of the first occurring is 0.01 and the chance of the second occurring is 0.02, then the chance of a train crash is the chance of both occurring together, or $0.01 \bullet 0.02 = 0.0002$. The same strategy has been used in nuclear reactors; as I type this I can look out my window and see a nuclear reactor across the Hudson River. For a serious nuclear accident to occur, three 'independent' systems must fail simultaneously: primary and secondary cooling systems plus the containment vessel. However, the resulting optimistic statements about reactor safety can be short-circuited by a single circumstance that prevents the independence of fail-safes (e.g., operator panic misjudgments or an earthquake).

Entire statistics books are written on probability, permitting calculation of probabilities for a wide suite of experimental conditions. Here our scope is much more humble: to consider a single 'probability distribution function' known as the normal distribution, and to determine how to assess the probabilities associated with a single variable or a relationship between two variables.

* * *

Sampling Distribution for One Variable

Begin with a variable which we will call *X*, for which we have 100 measurements. This dataset was drawn from a table of random normal numbers, but in the next section we will consider actual datasets of familiar data types. Usually we have minimal interest in the individual values of our 100 (or however many) measurements of variable *X*; these measurement values are simply a means to an

end. We are actually interested in knowing the true value of variable *X*, and we make replicate measurements in order to decrease the influence of random errors on our estimation of this true value. Using the term 'estimation' does not imply that one is guessing the value. Instead 'estimation' refers to the fact that measurements estimate the true value, but measurement errors of some type are almost always present.

Even if we were interested in the 100 individual values of *X*, we face the problem that a short or prolonged examination of a list of numbers provides minimal insight, because the human mind cannot easily comprehend a large quantity of numbers simultaneously. What we really care about is usually the essence or basic properties of the dataset, in particular:

• what is the average value?

• what is the scatter?

• are these data consistent with a theoretically predicted value for *X*?

• are these data related to another variable, *Y*?

With caution, each of the first three questions can be described with a single number, and that is the subject of this chapter. The engaging question of relationship to other variables is discussed in Chapter 3.

Histograms

The 100 measurements of *X* are more easily visualized in a histogram than in a tabulation of numbers. A histogram is a simple binning of the data into a suite of adjacent intervals of equal width. Usually one picks a fairly simple histogram range and interval increment. For example, our 100 measurements range from a minimum of -2.41 to a maximum of 2.20, so one might use a plot range of -2.5 to 2.5 and an interval of 0.5 or 1.0, depending on how many data points we have. The choice of interval is arbitrary but important, as it affects our ability to see patterns within the data. For example, Figure 1 shows that for the first 20 values of this dataset:

• an interval of 0.2 is too fine, because almost every data point goes into a separate bin. The eye tends to focus mainly on individual data points rather than on broad patterns, and we cannot easily see the relative frequencies of values.

• an interval of 0.5 or 1.0 is good, because we can see the overall pattern of a bell-shaped distribution without being too distracted by looking at each individual point.

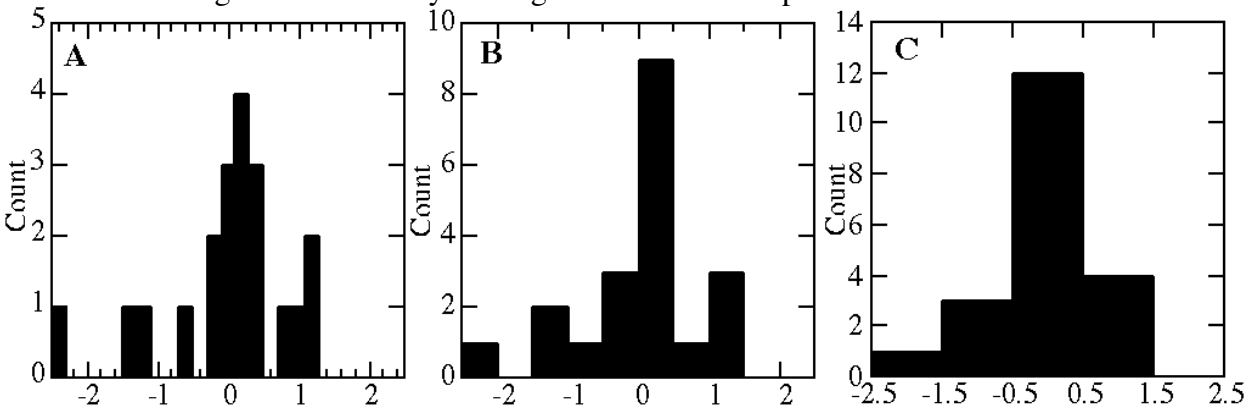


Figure 1. Three histograms of the same rand20a dataset, with binning intervals of 0.2 (A), 0.5 (B), and 1.0 (C). A longer binning interval helps to show that these data are from a normal distribution.

When we have 50 or 100 measurements instead of 20, we find that a finer histogram-binning interval is better for visualizing the pattern of the data. Figure 2 shows that an interval of about 0.5 is best for 100 measurements of this data type.

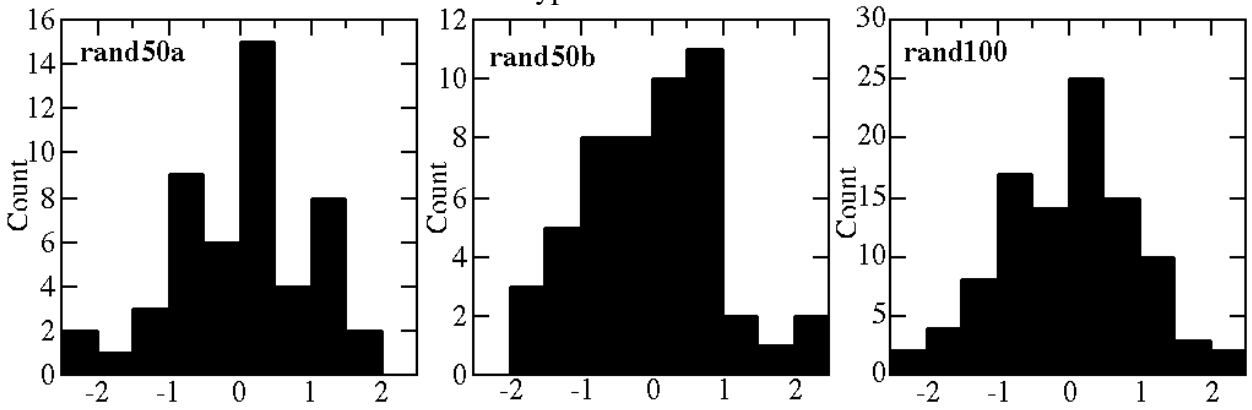


Figure 2. Histograms of two 50-point datasets (rand50a & rand50b) and a combined 100-point dataset (rand100). Although all data are drawn from a table of random normal numbers, rand50b appears to be non-normally distributed. Based on these histograms and Figure 1, a dataset must have more than 50 points for reliable visual determination of whether it is normally distributed.

* * *

Normal Distribution

The data shown in Figures 1 and 2 have what is called a **normal distribution**. Such a distribution is formally called a Gaussian distribution or informally called a bell curve. The normal distribution has both a theoretical and an empirical basis. Theoretically, we expect a normal distribution whenever some parameter or variable *X* has many independent, random causes of variation and several of these so-called 'sources of variance' have effects of similar magnitude. Even if an individual type of error is non-normally distributed, groups of such errors are. Empirically, countless types of measurements in all scientific fields exhibit a normal distribution. Yet *we must always verify the assumption that our data follow a normal distribution*. Failure to test this assumption is scientists' most frequent statistical pitfall. This mistake is needless, because one can readily examine a dataset to determine whether or not it is normally distributed.

Mean and Standard Deviation

For any dataset that follows a normal distribution, regardless of dataset size, virtually all of the information is captured by only three variables:

N: the number of data points, or measurements;

$\bar{X}$: the mean value; and

σ: the standard deviation.

The **mean** ($\bar{\mathbf{X}}$), also called the arithmetic mean, is an average appropriate only for normal distributions. The mean is defined as:

$$\bar{X} = \sum_{i=1}^{N} x_i /N \quad = (x_1+x_2+...x_{N-1}+x_N)/N$$

or, in shortened notation, $\bar{X} = \Sigma x_i/N$. The mean is simply the sum of all the individual measurement values, divided by the number of measurements.

The **standard deviation** (σ) is a measure of the dispersion or scatter of data. Defined as the square root of the **variance** (σ^2), it is appropriate only for normal distributions. The variance is defined as:

$\sigma^2 = \Sigma(x_i - \bar{X})^2/N$.

Thus the variance is the average *squared deviation* from the mean, i.e., the sum of squared deviations from the mean divided by the number of data points. Computer programs usually avoid handling each measurement twice (first to calculate the mean and later to calculate the variance) by using an alternative equation: $\sigma^2 = N^{-1}\Sigma(x_i^2) - \bar{X}^2$.

The standard deviation and variance are always positive. The units of standard deviation are the same as those of the *x* data. Often one needs to compare the scatter to the average value; two handy measures of this relationship are the fractional standard deviation ($\sigma/\bar{X}$) and percentage standard deviation ($100\sigma/\bar{X}$).

Normal Distribution Function

The **normal distribution function**, or 'normal error function', is shown in Figure 3. This probability distribution function of likely *X* values is expressed in terms of the 'true mean' *M* and standard deviation σ as:

$f(x) = (1/\sigma(2\pi)^{0.5})e^{-(x-M)^2/2\sigma^2}$.

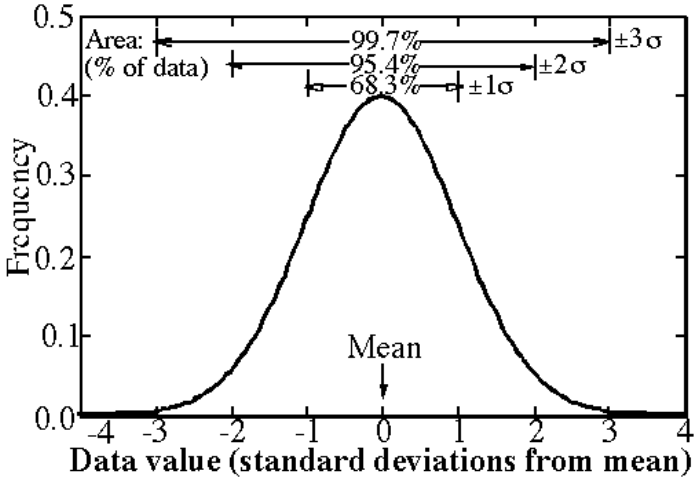


Figure 3. The normal distribution function.

For data drawn from a normal distribution, we can expect about 68.3% of the measurements to lie within one standard deviation of the mean, with half of the 68.3% above the mean and half below. Similarly, 95.4% of the measurements will lie within two standard deviations of the mean (i.e., within the interval $\bar{X}-2\sigma < x_i < \bar{X}+2\sigma$), and 99.7% of the measurements will lie within three standard deviations of the mean. These percentages are the areas under portions of the normal distribution function, as shown in Figure 3. All statistics books explain how to find the area under any desired portion of the curve, i.e., how to find the expected proportion of the data that will have values between specified limits. Of course, for the finite number of measurements of an individual dataset, we will only approximately observe these percentages. Nevertheless, it is well worth memorizing the following two ap-

proximations: *two thirds of data lie within one standard deviation of the mean, and 95% lie within two standard deviations*.

Although calculation of both the mean and standard deviation involves division by N, both are relatively independent of N. In other words, increasing the number of data points does not systematically increase or decrease either the mean or the standard deviation. Increasing the number of data points, however, does increase the usefulness of our calculated mean and standard deviation, because it increases the reliability of inferences drawn from them.

Based on visual examination of a histogram, it may be difficult to tell whether or not the data originate from a normally distributed parent population. For small N such as N=20 in Figure 1, random variations can cause substantial departures from the bell curve of Figure 3, and only the coarsest binning interval (Figure 1c) looks somewhat like a simple normal distribution. Even with N=50, the two samples in Figure 2 visually appear to be quite different, although both were drawn from the same table of random normal numbers. With N=100, the distribution begins to approximate closely the theoretical normal distribution (Figure 3) from which it was drawn. Fortunately, the mean and standard deviation are more robust; they are very similar for the samples of Figures 1 and 2.

The mean provides a much better estimate of the true value of X (the 'true mean' M) than do any of the individual measurements of X, because the mean averages out most of the random errors that cause differences between the individual measurements. How much better the mean is than the individual measurements depends on the dispersion (as represented by the standard deviation) and the number of measurements (N); more measurements and smaller standard deviation lead to greater accuracy of the calculated mean in estimating the true mean.

Our sample of N measurements is a subset of the parent population of potential measurements of X. We seek the value M of the parent population (the 'true mean'). Finding the average $\bar{X}$ of our set of measurements (the 'calculated mean') is merely a means to that end. We are least interested in the value x_i of any individual measurement, because it is affected strongly by unknown and extraneous sources of noise or scatter. *If the data are normally distributed and unbiased, then the calculated mean is the most probable value of the true average of the parent population*. Thus the mean is an extremely important quantity to calculate. Of course, if the data are biased such as would occur with a distorted yardstick, then our estimate of the true average is also biased. We will return later to the effects of a non-normal distribution.

Just as one can determine the mean and standard deviation of a set of N measurements, one can imagine undertaking several groups of N measurements and then calculating a grand mean and standard deviation of these groups. This grand mean would be closer to the true mean than most of the individual means would be, and the scatter of the several group means would be smaller than the scatter of the individual measurements. The standard deviation of the mean ($\sigma_{\bar{x}}$), also called the **standard error of the mean**, is: $\boldsymbol{\sigma_{\bar{x}}=\sigma/N^{0.5}}$. Note that unlike a sample standard deviation, the standard deviation of the mean *does* decrease with increasing N. This standard deviation of the mean has three far-reaching but underutilized applications: weighted averages, confidence limits for the true mean, and determining how many measurements one should make.

Weighted Mean

A weighted mean is the best way to average data that have different precisions, if we know or can estimate those precisions. The weighted mean is calculated like an ordinary mean, except that we multiply each measurement by a weighting factor and divide the sum of these products not by *N* but by the sum of the weights, as follows:

$\bar{X} = \Sigma w_i x_i / \Sigma w_i$ where w_i is the weighting factor of the *i*th measurement. If we use equal weights, then this equation reduces to the equation for the ordinary mean. Various techniques for weighting can be used. If each of the values to be averaged is itself a mean with an associated known variance, then the most theoretically satisfying procedure is to weight each value according to the inverse of the variance of the mean: $w_i = 1/\sigma^2_{\bar{x}_i} = N/\sigma^2_i$. The weighted variance is: $\sigma^2_{\bar{x}} = 1/\Sigma(1/\sigma^2_{\bar{x}_i}) = 1/\Sigma w_i$.

For example, suppose that three laboratories measure a variable *Y* and obtain the following:

	N	mean	σ	$\sigma_{\bar{x}}$ (=σN$^{-0.5}$)	w_i
lab 1:	20	109	10	2.24	0.2
lab 2:	20	105	7	1.57	0.41
lab 3:	50	103	7	0.99	1.02

Then $\bar{X}$ =(0.20•109 + 0.41•105 + 1.02•103)/(0.20+0.41+1.02) = 104.2. The variance of this weighted mean is $\sigma^2_{\bar{x}} = 1/(0.20+0.41+1.02) = 0.613$, and so the standard deviation of the weighted mean is $\sigma_{\bar{x}} = 0.78$. Note that the importance or weighting of the measurements from Lab 2 is twice as high as from Lab 1, entirely because Lab 2 was able to achieve a 30% lower standard deviation of measurements than Lab 1 could. Lab 3, which obtained the same standard deviation as Lab 2 but made 2.5 times as many measurements as Lab 2, has 2.5 times the importance or weighting of results.

95% Confidence Limits on Mean

Usually we want to use our measurements to make a quantitative estimate of the true mean *M*. One valuable way of doing so is to state the 95% confidence limits on the true mean, which for convenience we will call α_{95}. Confidence limits for the true mean *M* can be calculated as follows:

95% confidence limits: $\alpha_{95}=\sigma_{\bar{x}}\bullet t_{95}$ $\quad \bar{X}-\alpha_{95} < M < \bar{X}+\alpha_{95}$

99% confidence limits: $\alpha_{99}=\sigma_{\bar{x}}\bullet t_{99}$ $\quad \bar{X}-\alpha_{99} < M < \bar{X}+\alpha_{99}$

Just multiply the standard error of the mean by the 't-factor', finding the t-factor in the table below for the appropriate number of measurements.

By stating the mean (our best estimate of the true mean *M*) and its 95% confidence, we are saying that there is only a 5% chance that the true mean is outside the range $\bar{X}\pm\alpha_{95}$. One's desire to state results with as high a confidence level as possible is countered by the constraint that higher confidence levels encompass much broader ranges of potential data values. For example, our random-number dataset (N=100, $\sigma_{\bar{x}}$=0.095, $\bar{X}$=0.02) allows us to state with 95% confidence that the true mean lies within the interval -0.17 to 0.21 (i.e., $\bar{X} \pm \alpha_{95}$, or 0.02 ± 0.19). We can state with

99% confidence that the true mean is within the interval -0.23 to 0.27. Actually the true mean for this dataset is zero.

Table 1. Values of the *t* distribution for 95% and 99% confidence limits (two-tailed) and for different sample sizes [Fisher and Yates, 1963].

N:	2	3	4	5	6	7	8	9	10	11
t_{95}:	12.71	4.303	3.182	2.776	2.571	2.447	2.365	2.306	2.262	2.228
t_{99}:	63.66	9.925	5.841	4.604	4.032	3.707	3.499	3.355	3.250	3.169
N:	12	13	14	15	16	17	18	19	20	21
t_{95}:	2.201	2.179	2.160	2.145	2.131	2.120	2.110	2.101	2.093	2.086
t_{99}:	3.106	3.055	3.012	2.977	2.947	2.921	2.898	2.878	2.861	2.845
N:	22	23	24	25	30	40	60	80	100	∞
t_{95}:	2.080	2.074	2.069	2.064	2.045	2.023	2.001	1.990	1.984	1.960
t_{99}:	2.831	2.819	2.807	2.797	2.756	2.713	2.662	2.640	2.627	2.576

Selection of a confidence level (α_{95}, α_{99}, etc.) usually depends on one's evaluation of which risk is worse: the risk of incorrectly identifying a variable or effect as significant, or the risk of missing a real effect. Is the penalty for error as minor as having a subsequent researcher correct the error, or could it cause disaster such as an airplane crash? If prior knowledge suggests one outcome for an experiment, then rejection of that outcome needs a higher than ordinary confidence level. For example, no one would take seriously a claim that an experiment demonstrates test-tube cold fusion at the 95% confidence level; a much higher confidence level plus replication was demanded. Most experimenters use either a 95% or 99% confidence level. Tables for calculation of confidence limits other than 95% or 99%, called tables of the *t* distribution, can be found in any statistics book.

How Many Measurements are Needed?

The standard error of the mean $\sigma_{\bar{x}}$ is also the key to estimating how many measurements to make. The definition $\sigma_{\bar{x}}=\sigma N^{-0.5}$ can be recast as $N=\sigma^2/\sigma^2_{\bar{x}}$. Suppose we want to make enough measurements to obtain a final mean that is within 2 units of the true mean (i.e., $\sigma_{\bar{x}}\leq 2$), and a small pilot study permits us to calculate that our measurement scatter $\sigma\approx 10$. Then our experimental series will need $N\geq 10^2/2^2$, or $N\geq 25$, measurements to obtain the desired accuracy at the 68% confidence level (or $1\sigma_{\bar{x}}$). For about 95% confidence, we recall that about 95% of points are within 2σ of the mean and conclude that we would need $2\sigma_{\bar{x}}\leq 2$, so $N\geq 10^2/1^2$, or $N\geq 100$ measurements. Alternatively and more accurately, we can use the *t* table above to determine how many measurements will be needed to assure that our mean is within 2 units of the true mean at the 95% confidence level ($\alpha_{95}\leq 2$): we need for $t_{95}=\alpha_{95}/\sigma_{\bar{x}}=\alpha_{95}\, N^{0.5}/\sigma = 2\, N^{0.5}/10 = 0.2\, N^{0.5}$ to be greater than the t_{95} in the table above for that *N*. By trying a few values of *N*, we see that $N\geq 100$ is needed.

As a rule of thumb, *one must quadruple the number of measurements in order to double the precision of the result*. This generalization is based on the $N^{0.5}$ relationship of standard deviation to standard error and is strictly true only if our measure of precision is the standard error. If, as is of-

ten the case, our measure of precision is α_{95}, then the rule of thumb is only approximately true because the *t*'s of Table 1 are only approximately equal to 2.0.

* * *

Propagation of Errors

Sometimes the variable of interest actually is calculated from measurements of one or more other variables. In such cases it is valuable to see how errors in the measured variables will propagate through the calculation and affect the final result. Propagation of errors is a scientific concern for several reasons:

- it permits us to calculate the uncertainty in our determination of the variable of interest;
- it shows us the origin of that uncertainty; and
- a quick analysis of propagation of errors often will tell us where to concentrate most of our limited time resources.

If several different independent errors (e_i) are responsible for the total error (E) of a measurement, then:

$$E^2 = e_1^2+e_2^2+\ldots+e_N^2$$

As a rule of thumb, one can *ignore any random error that is less than a quarter the size of the dominant error*. The squaring of errors causes the smaller errors to contribute trivially to the total error. If we can express errors in terms of standard deviations and if we have a known relationship between error-containing variables, then we can replace the estimate above with the much more powerful analysis of propagation of errors which follows.

Suppose that the variable of interest is V, and it is a function of the several variables $a, b, c, \ldots$: $V=f(a,b,c,...)$. If we know the variances of $a, b, c, \ldots$, then the variance of V can be calculated from:

$$\sigma^2_V = (\partial V/\partial a)^2 \bullet \sigma^2_a + (\partial V/\partial b)^2 \bullet \sigma^2_b + \ldots \qquad (1)$$

Thus the variance of V is equal to the sum of the product of each individual variance times the square of the partial derivative. For example, if we want to determine the area (A) of a rectangle by measuring its two sides (a and b): $A=ab$, and $\sigma^2_A = (\partial A/\partial a)^2 \bullet \sigma^2_a + (\partial A/\partial b)^2 \bullet \sigma^2_b = \bar{b}^2\sigma^2_a + \bar{a}^2\sigma^2_b$. Propagation of errors can be useful even for single-variable problems. For example, if we want to determine the area (A) of a circle by measuring its radius (r): $A=\pi r^2$, and $\sigma^2_A = (\partial A/\partial r)^2 \bullet \sigma^2_r = (2\pi\ \bar{r})^2\sigma^2_r$.

Why analyze propagation of errors? In the example above of determining area of a circle from radius, we could ignore propagation of errors, just convert each radius measurement into an area, and then calculate the mean and standard deviation of these area determinations. Similarly, we could calculate rectangle areas from pairs of measurements of sides a and b, then calculate the mean and standard deviation of these area determinations. In contrast, each of the following variants on the rectangle example would benefit from analyzing propagation of errors:

- measurements a and b of the rectangle sides are not paired; shall we arbitrarily create pairs for calculation of A, or use propagation of errors?
- we have different numbers of measurements of rectangle sides a and b. We must either discard some measurements or, better, use propagation of errors;

• we are about to measure rectangle sides *a* and *b* and we know that *a* will be about 10 times as big as *b*. Because $\sigma^2_A = \bar{b}^2\sigma^2_a + \bar{a}^2\sigma^2_b$, the second term will be about 100 times as important as the first term if *a* and *b* have similar standard deviations, and we can conclude that it is much more important to find a way to reduce σ^2_b than to reduce σ^2_a.

Usually we are less interested in the variance of *V* than in the variance of the mean *V*, or its square root (the standard error of *V*). We can simply replace the variances in equation (1) above with variances of means. Using variances of means, propagation of errors allows us to estimate how many measurements of each of the variables *a,b,c,. . .* would be needed to determine *V* with some desired level of accuracy, if we have a rough idea of what the expected variances of *a,b,c, . . .* will be. Typically the variables *a,b,c,. . .* will have different variances which we can roughly predict after a brief pilot study or before we even start the controlled measurement series. If so, *a quick analysis of propagation of errors will suggest concentrating most of our limited time resources on one variable*, either with a large number of measurements or with slower and more accurate measurements. For example, above we imagined that *a* is about 10 times as big as *b* and therefore concluded that we should focus on reducing σ^2_b instead of reducing σ^2_a. Even if we have no way of reducing σ^2_b, we can reduce $\sigma^2_{\bar{b}}$(variance of mean b) by increasing the number of measurements, because the standard error $\sigma_{\bar{x}}=\sigma N^{-0.5}$.

Equation (1) and ability to calculate simple partial derivatives will allow one to analyze propagation of errors for most problems. Some problems are easier if equation (1) is recast in terms of fractional standard deviations:

$$(\sigma_v/V)^2 = (V^{-1}\bullet\partial V/\partial a)^2\bullet\sigma^2_a + (V^{-1}\bullet\partial V/\partial b)^2\bullet\sigma^2_b + \ldots \qquad (2)$$

Based on equation (1) or (2), here are the propagation of error equations for several common relationships of *V* to the variables *a* and *b*, where *k* and *n* are constants:

V=ka+nb: $\sigma^2_v = k^2\sigma^2_a + n^2\sigma^2_b$

V=ka-nb: $\sigma^2_v = k^2\sigma^2_a + n^2\sigma^2_b$

V=kab: $\sigma^2_v = (k\ \bar{b}\sigma_a)^2 + (k\ \bar{a}\sigma_b)^2$

or: $(\sigma_v/\ \bar{V})^2 = (\sigma_a/\ \bar{a})^2 + (\sigma_b/\ \bar{b})^2$

V=ka/b: $(\sigma_v/\ \bar{V})^2 = (\sigma_a/\ \bar{a})^2 + (\sigma_b/\ \bar{b})^2$

V=ka^n: $\sigma_v/\ \bar{V} = n\sigma_a/\ \bar{a}$

V=a^{k}b^n: $(\sigma_v/\ \bar{V})^2 = (k\sigma_a/\ \bar{a})^2 + (n\sigma_b/\ \bar{b})^2$

* * *

Non-Normal Distributions

The most frequent statistics pitfall is also a readily avoided *pitfall: assuming a normal distribution when the data are non-normally distributed*. Every relationship and equation in the previous section should be used only if the data are normally distributed or at least approximately normally distributed. The more data depart from a normal distribution, the more likely it is that one will be

misled by using what are called 'parametric statistics', i.e., statistics that assume a Gaussian distribution of errors. This section is organized in the same sequence that most data analyses should follow:

1) test the data for normality;

2) if non-normal, can one transform the data to make them normal?

3) if non-normal, should anomalous points be omitted?

4) if still non-normal, use non-parametric statistics.

Normality Tests

Because our statistical conclusions are often somewhat dependent on the assumption of a normal distribution, we would like to undertake a test that permits us to say "I am 95% confident that this distribution is normal." But such a statement is no more possible than saying that we are 95% certain that a hypothesis is correct; disproof is more feasible and customary than proof. Thus our normality tests may allow us to say that "there is <5% chance that this distribution is normal" or, in statistical jargon, "We reject the null hypothesis of a normal distribution at the 95% confidence level."

Experienced scientists usually test data for normality subjectively, simply by looking at a histogram and deciding that the data look approximately normally distributed. Yet I, an experienced scientist, would not have correctly interpreted the center histogram of Figure 2 as from a normal distribution. If in doubt, one can apply statistical tests of normality such as Chi-square (χ^2) and examine the type of departure from normality with measures such as skewness. Too often, however, even the initial subjective examination is skipped.

We can use a **χ^2 test** to determine whether or not our data distribution departs substantially from normality. A detailed discussion of the many applications of χ^2 tests is beyond the scope of this book, but almost all statistics books explain how a χ^2 test can be used to compare any data distribution to any theoretical distribution. A χ^2 test is most easily understood as a comparison of a data histogram with the theoretical Gaussian distribution. The theoretical distribution predicts how many of our measurements are expected to fall into each histogram bin. Of course, this expected frequency [Nf(n)] for the nth bin (or interval) will differ somewhat from the actual data frequency [F(n)], or number of values observed in that interval. Indeed, we saw in Figure 2 that two groups of 50 normally distributed measurements exhibited surprisingly large differences both from each other and from the Gaussian distribution curve. The key question then is how much of a difference between observed frequency and predicted frequency is chance likely to produce. The variable χ^2, which is a measure of the goodness of fit between data and theory, is the sum of squares of the fractional differences between expected and observed frequencies in all of the histogram bins:

$$\chi^2 = \sum_{n}\{[\mathrm{Nf}(n)-\mathrm{F}(n)]^2/\mathrm{Nf}(n)\} \qquad (3)$$

Comparison of the value of χ^2 to a table of predicted values allows one to determine whether statistically significant non-normality has been detected. The table tells us the range of χ^2 values that are typically found for normal distributions. We do not expect values very close to zero, indi-

cating a perfect match of data to theory. Nor do we expect χ^2 values that are extremely large, indicating a huge mismatch between the observed and predicted distributions.

The χ^2 test, like a histogram, can use any data units and almost any binning interval, with the same proviso that a fine binning interval is most appropriate when N is large. Yet some χ^2 tests are much easier than others, because of the need to calculate a predicted number of points for each interval. Here we will take the preliminary step of **standardizing** the data. Standardization transforms each measurement x_i into a unitless measurement which we will call z_i, where $\mathbf{z_i = (x_i - \bar{X})/\sigma}$. Standardized data have a mean of zero and a standard deviation of one, and any standardized array of approximately normally distributed data can be plotted on the same histogram. If we use a binning interval of 0.5σ, then the following table of areas under a normal distribution gives us the expected frequency [Nf(***n***)=N•area] in each interval.

Table 2: Areas of intervals of the normal distribution [Dixon and Massey, 1969].

σ Interval:	<-3	-3 to -2.5	-2.5 to -2	-2 to -1.5	-1.5 to -1	-1 to -0.5	-0.5 to 0.0
Area:	0.0013	0.0049	0.0166	0.044	0.0919	0.1498	0.1915
σ Interval:	>3	3 to 2.5	2.5 to 2	2 to 1.5	1.5 to 1	1 to 0.5	0.5 to 0.0

Equation 3 is applied to these 14 intervals, comparing the expected frequencies to the observed frequencies of the standardized data. Note that the intervals can be of unequal width. If the number of data points is small (e.g., *N*<20), one should reduce the 14 intervals (***n***=14) to 8 intervals by combining adjacent intervals of Table 2 [e.g., *f(**n**)* for 2σ to 3σ is .0166+.0049=.0215]. The following table shows the probabilities of obtaining a value of χ^2 larger than the indicated amounts, for ***n***=14 or ***n***=8. Most statistics books have much more extensive tables of χ^2 values for a variety of 'degrees of freedom' (df). When using such tables to compare a sample distribution to a Gaussian distribution that is estimated from the data rather than known independently, then df=***n***-2 as in Table 3.

Table 3. Maximum values of χ^2 that are expected from a normal distribution for different numbers of binning intervals (***n***) at various probability levels (P) [Fisher and Yates, 1963].

	P_{80}	P_{90}	P_{95}	$P_{97.5}$	P_{99}	$P_{99.5}$
n=8:	8.56	10.64	12.59	14.45	16.81	18.55
n=14:	15.81	18.55	21.03	23.34	26.22	28.3

For example, for ***n***=14 intervals a χ^2 value of 22 (calculated from equation 3) would allow one to reject the hypothesis of a normal distribution at the 95% confidence level but not at 97.5% confidence (21.03<22<23.34).

A non-normal value for χ^2 can result from a single histogram bin that has an immense difference between predicted and observed value; it can also result from a consistent pattern of relatively small differences between predicted and observed values. Thus the χ^2 test only determines whether, not how, the distribution may differ from a normal distribution.

Skewness is a measure of how symmetric the data distribution is about its mean. A distribution is positively skewed, or skewed to the right, if data extend substantially farther to the right of the peak than they do the left. Conversely, a distribution is negatively skewed, if data extend substan-

tially farther to the left of the peak. A normal distribution is symmetric and has a skewness of zero. Later in this chapter we will see several examples of skewed distributions. A rule of thumb is that *the distribution is reasonably symmetric if the skewness is between -0.5 and 0.5, and the distribution is highly skewed if the skewness is <-1 or >1.*

* * *

If a data distribution is definitely non-normal, it might still be possible to transform the dataset into one that is normally distributed. Such a transformation is worthwhile, because it permits use of the parametric statistics above, and we shall soon see that parametric statistics are more efficient than non-parametric statistics. In some fields, transformations are so standard that the ordinary untransformed mean is called the arithmetic mean to distinguish it from means based on transformations.

The most pervasively suitable transformation is logarithmic: either take the natural logarithm of all measurements and then analyze them using techniques above, or simply calculate the **geometric mean** ($\bar{g}$): $\bar{g} = \Sigma(x_i)^{1/N}$. The geometric mean is appropriate for ratio data and data whose errors are a percentage of the average value. If data are positively skewed, it is worth taking their logarithms and redoing the histogram to see if they look more normal. More rarely, normality can be achieved by taking the inverse of each data point or by calculating a **harmonic mean** ($\bar{h}$): $\bar{h} = N/\Sigma(1/x_i)$.

* * *

Rejecting Anomalous Data

Occasionally a dataset has one or more anomalous data points, and the researcher is faced with the difficult decision of rejecting anomalous data. In Chapter 6, we consider the potential pitfalls of rejecting anomalous data. In many scientists' minds, data rejection is an ethical question: some routinely discard anomalous points without even mentioning this deletion in their publication, while others refuse to reject any point ever. Most scientists lie between these two extremes.

My own approach is the following:

- publish all data,
- flag points that I think are misleading or anomalous and explain why I think they are anomalous,
- show results either without the anomalous points or both with and without them, depending on how confident I am that they should be rejected.

In this way I allow the reader to decide whether rejection is justified, and the reader who may wish to analyze the data differently has *all* of the data available. Remembering that sometimes anomalies are the launching point for new insights, no scientist should hide omitted data from readers.

Here we will consider the question of data rejection statistically: are there statistical grounds for rejecting a data point? For example, if we have 20 measurements, we can expect about one measurement to differ from the mean by more than 2σ, but we expect (Table 2) that only 0.13% of the data points will lie more than three standard deviations below the mean. If one point out of 20 differs from the mean by more than 3σ, we can say that such an extreme value is highly unlikely to occur by chance as part of the same distribution function as the other data. Effectively, we are deciding that this anomalous point was affected by an unknown different variable. Can we conclude therefore that it should be rejected?

Although the entire subject of data rejection is controversial, an objective rejection criterion seems preferable to a subjective decision. One objective rejection criterion is **Chauvenet's criterion**: a measurement can be rejected if the probability of obtaining it is less than $1/2N$. For example, if N=20 then a measurement must be so distant from the mean that the probability of obtaining such a value is less than 1/40 or 2.5%. Table 4 gives these cutoffs, expressed as the ratio of the observed deviation (d_i) to the standard deviation, where the deviation from the mean is simply $d_i = |x_i - \bar{X}|$.

Table 4. Deviation from the mean required for exclusion of a data point according to Chauvenet's criterion [Young, 1962].

N:	5	6	7	8	9	10	12	14	16	18	20
d_i/σ:	1.65	1.73	1.81	1.86	1.91	1.96	2.04	2.1	2.15	2.2	2.24
N:	25	30	40	50	60	80	100	150	200	400	1000
d_i/σ:	2.33	2.39	2.49	2.57	2.64	2.74	2.81	2.93	3.02	3.23	3.48

What mean and standard deviation should one use in applying Chauvenet's criterion? The calculated mean and especially standard deviation are extremely sensitive to extreme points. Including the suspect point in the calculation of $\bar{X}$ and σ substantially decreases the size of d_i/σ and thereby decreases the likelihood of rejecting the point. Excluding the suspect point in calculating the mean and standard deviation, however, is tantamount to assuming *a priori* what we are setting out to test; such a procedure often would allow us to reject extreme values that are legitimate parts of the sample population. Thus we must take the more conservative approach: the mean and standard deviation used in applying Chauvenet's criterion should be those calculated *including* the suspect point.

If Chauvenet's criterion suggests rejection of the point, then the final mean and standard deviation should be calculated excluding that point. In theory, one then could apply the criterion again, possibly reject another point, recalculate mean and standard deviation again, and continue until no more points can be rejected. In practice, this exclusion technique should be used sparingly, and applying it more than once to a single dataset is not recommended.

Often one waffles about whether or not to reject a data point even if rejection is permitted by Chauvenet's criterion. Such doubts are warranted, for we shall see in later examples that Chauvenet's criterion occasionally permits rejection of data that are independently known to be reliable. An alternative to data rejection is to use some of the nonparametric statistics of the next section, for they are much less sensitive than parametric techniques are to extreme values.

* * *

Median, Range, and 95% Confidence Limits

Until now, we have used parametric statistics, which assume a normal distribution. Nonparametric statistics, in contrast, make no assumption about the distribution. Most scientific studies employ parametric, not nonparametric, statistics, for one of four reasons:

- experimenter ignorance that parametric statistics should only be applied to normal distributions;
- lack of attention to whether or not one's data are normally distributed;
- ignorance about nonparametric statistical techniques;
- greater efficiency of parametric statistics.

The first three reasons are inexcusable; only the last reason is scientifically valid. In statistics, as in any other field, *assumptions decrease the scope of possibilities and enable one to draw conclusions with greater confidence, if the assumption is valid*. For example, various nonparametric techniques require 5-50% more measurements than parametric techniques need to achieve the same level of confidence in conclusions. Thus nonparametric techniques are said to be less efficient than parametric techniques, and the latter are preferable if the assumption of a normal distribution is valid. If this assumption is invalid but made anyway, then parametric techniques not only overestimate the confidence of conclusions but also give somewhat biased estimates.

The nonparametric analogues of parametric techniques are:

Measure	Parametric	Nonparametric
Average:	Mean	Median
Dispersion:	Standard deviation	Interquartile range
Confidence limits:	Conf. limits on mean	Conf. limits on median

Nonparametric statistics are easy to use, whether or not they are an option in one's spreadsheet or graphics program. The first step in nearly all nonparametric techniques is to sort the measurements into increasing order. This step is a bit time consuming to do by hand for large datasets, but today most datasets are on the computer, and many software packages include a 'sort' command. We will use the symbol I_i to refer to the data value in sorted array position i; for example, I_1 would be the smallest data value.

The nonparametric measure of the true average value of the parent population is the **median**. For an odd number of measurements, the median is simply the middle measurement ($I_{N/2}$), i.e., that measurement for which half of the other measurements is larger and half is smaller. For an even number of measurements there is no single middle measurement, so the median is the average (midpoint) of the two measurements that bracket the middle. For example, if a sorted dataset of five points is 2.1, 2.1, 3.4, 3.6, and 4.7, then the median is 3.4; if a sorted dataset of six points is 2.1, 2.1, 3.4, 3.6, 4.7, and 5.2, then the median is (3.4+3.6)/2 = 3.5.

The median divides the data population at the 50% level: 50% are larger and 50% are smaller. One can also divide a ranked dataset into four equally sized groups, or quartiles. One quarter of the data are smaller than the first quartile, the median is the second quartile, and one quarter of the data are larger than the third quartile.

The **range** is a frequently used nonparametric measure of data dispersion. The range is the data pair of smallest (I_1) and largest (I_N) values. For example, the range of the latter dataset above is 2.1-5.2. The range is a very inefficient measure of data dispersion; one measurement can change it dramatically. A more robust measure of dispersion is the interquartile range, the difference between the third and first quartiles. The interquartile range ignores extreme values. It is conceptually analogous to the standard deviation: the interquartile range encompasses the central 50% of the data, and ±1 standard deviation encompasses the central 68% of a normal distribution,

For non-normal distributions, **confidence limits for the median** are the best way to express the reliability with which the true average of the parent population can be estimated. Confidence limits are determined by finding the positions I_k and I_{N-k+1} in the sorted data array I_i, where k is determined from Table 5 below. Because these confidence limits use an integer number of array positions, they do not correspond exactly to 95% or 99% confidence limits. Therefore Table 5 gives the largest k yielding a probability of at least the desired probability. For example, suppose that we have 9 ranked measurements: 4.5, 4.6, 4.9, 4.9, 5.2, 5.4, 5.7, 5.8, and 6.2. Then N=9, k=3 yields less

than 95% confidence, *k*=2 yields the 96.1% confidence limits 4.6-5.8, and *k*=1 yields 99.6% confidence limits 4.5-6.2.

Table 5. Confidence limits for the median [Nair, 1940; cited by Dixon and Massey, 1969].

N	k	α>95	k	α>99	N	k	α>95	k	α>99	N	k	α>95	k	α>99
6	1	96.9	-		26	8	97.1	7	99.1	46	16	97.4	14	99.5
7	1	98.1	-		27	8	98.1	7	99.4	47	17	96	15	99.2
8	1	99.2	1	99.2	28	9	96.4	7	99.6	48	17	97.1	15	99.4
9	2	96.1	1	99.6	29	9	97.6	8	99.2	49	18	95.6	16	99.1
10	2	97.9	1	99.8	30	10	95.7	8	99.5	50	18	96.7	16	99.3
11	2	98.8	1	99.9	31	10	97.1	8	99.7	51	19	95.1	16	99.5
12	3	96.1	2	99.4	32	10	98	9	99.3	52	19	96.4	17	99.2
13	3	97.8	2	99.7	33	11	96.5	9	99.5	53	19	97.3	17	99.5
14	3	98.7	2	99.8	34	11	97.6	10	99.1	54	20	96	18	99.1
15	4	96.5	3	99.3	35	12	95.9	10	99.4	55	20	97	18	99.4
16	4	97.9	3	99.6	36	12	97.1	10	99.6	56	21	95.6	18	99.5
17	5	95.1	3	99.8	37	13	95.3	11	99.2	57	21	96.7	19	99.2
18	5	96.9	4	99.2	38	13	96.6	11	99.5	58	22	95.2	19	99.5
19	5	98.1	4	99.6	39	13	97.6	12	99.1	59	22	96.4	20	99.1
20	6	95.9	4	99.7	40	14	96.2	12	99.4	60	22	97.3	20	99.4
21	6	97.3	5	99.3	41	14	97.2	12	99.6	61	23	96	21	99
22	6	98.3	5	99.6	42	15	95,6	13	99.2	62	23	97	21	99.3
23	7	96.5	5	99.7	43	15	96.8	13	99.5	63	24	95.7	21	99.5
24	7	97.7	6	99.3	44	16	95.1	14	99	64	24	96.7	22	99.2
25	8	95.7	6	99.6	45	16	96.4	14	99.3	65	25	95.4	22	99.4

Nonparametric statistics make no assumptions about the shape of either the parent population or the data distribution function. Thus nonparametric statistics cannot recognize that any data value is anomalous, and data rejection criteria such as Chauvenet's criterion are impossible. In a sense, nonparametric statistics are intermediate between rejection of a suspect point and blind application of parametric statistics to the entire dataset; no points are rejected, but the extreme points receive much less weighting than they do when a normal distribution is assumed.

One fast qualitative ('quick-and-dirty') test of the suitability of parametric statistics for one's dataset is to see how similar the mean and median are. If the difference between them is minor in comparison to the size of the standard deviation, then the mean is probably a reasonably good estimate, unbiased by either extreme data values or a strongly non-normal distribution. A rule of thumb might be to suspect non-normality or anomalous extreme values if $4(\bar{X}-\tilde{X})>\sigma$, where $\tilde{X}$ is the median.

* * *

Figure 4 is a flowchart that shows one possible way of approaching analysis of a variable. Rarely does anyone evaluate a variable as systematically as is shown in Figure 4; indeed, I have never seen such a flowchart or list of steps. This flowchart demonstrates why different examples, such as those in the following section, require different treatments.

A useful first step in analyzing a variable is to ask oneself whether the individual observations, measurements, or data are independent. Two events are **independent** if they are no more likely to be similar than any two randomly selected members of the population. Independence is implicit in the idea of random errors; with random errors we expect that adjacent measurements in our dataset will be no more similar to each other than distant measurements (e.g., first and last measurements) will be. Independence is an often-violated assumption of the single-variable statistical techniques. Relaxation of this assumption sometimes is necessary and permissible, as long as we are aware of the possible complications introduced by this violation (note that most scientists would accept this statement pragmatically, although to a statistician this statement is as absurd as saying A≠A). Except for the random-number example, none of the example datasets to follow has truly independent samples. We will see that lack of independence is more obvious for some datasets than for others, both in *a priori* expectation and in data analysis.

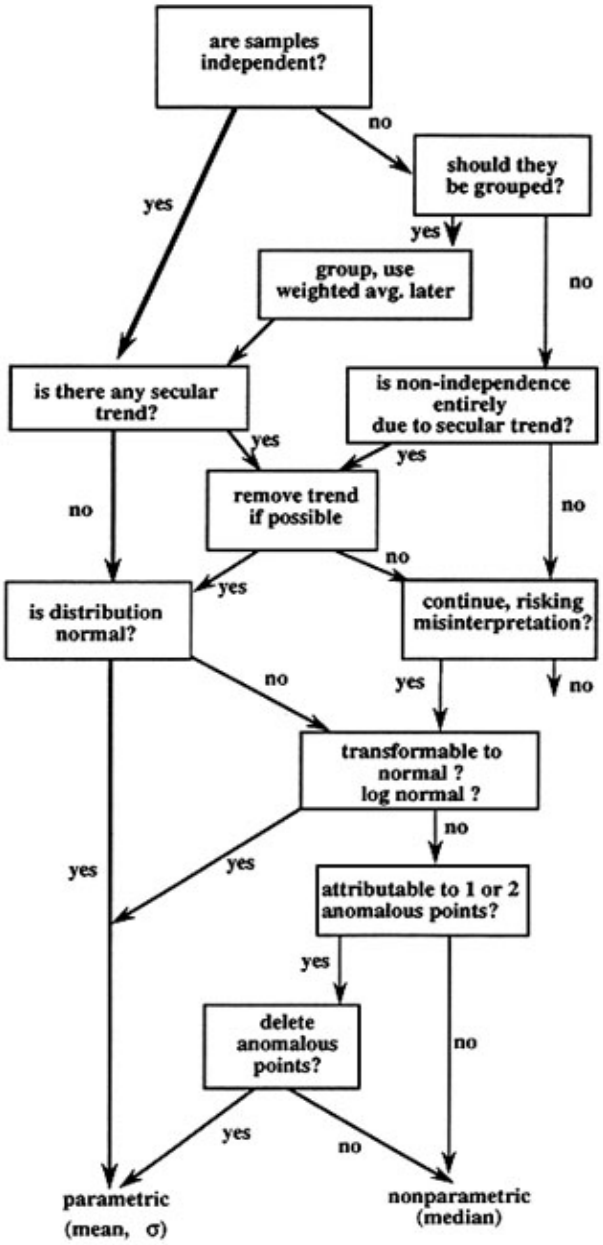


Figure 4. Flowchart of decision steps for a group of measurements.

Actual scientific data have the same problem: sometimes we expect our measurements to be unavoidably non-independent, whereas at other times we expect independence but our analysis reveals non-independence. Thus, regardless of expectations, *one should plot every dataset as a function of measurement sequence*, for visual detection of any unexpected secular trends. Examination of the data table itself often is an inadequate substitute. No statistical test detects secular trends as consistently as simple examination of a plot of variable vs. measurement order. Examples of such unexpected secular trends are:

- instrumental drift;
- measurement error during part of the data acquisition;

- undetected, involuntary change of part of the measurement procedure during the measurement series;
- undetected change in standards;
- temporal change in an unidentified relevant variable, i.e., a source of 'noise'.

* * *

Examples

We can gain insight into the statistical techniques described above by considering their application to a few datasets of different types. Our examples come from a variety of original sources, but I got almost all of them from the World Almanac [Hoffman, 1990]. The exceptions are the examples of random normal numbers and of the hare and tortoise. I have deliberately chosen familiar examples rather than artificial data or real scientific data, because the explanation for observed statistical behavior is easier to comprehend with familiar examples. The examples are:

- reexamination of the random normal numbers of Figures 1 and 2;
- race between the hare and the tortoise;
- percentage of high school students that graduate, by state;
- state population (1990 census);
- state taxes, per capita, by state;

Table 6 summarizes the statistical results for these examples, as well as some examples introduced in the next chapter.

Table 6. Summary statistics for the example problems used in this chapter and in Chapter 3. Statistics for population, taxes, and batting averages are shown both before and after exclusion of extreme points. Columns 2-7: parametric; columns 8-10: nonparametric; column 11: exclusion by Chauvenet's criterion (Y or N).

dataset	N	$\bar{X}$	σ	α_{95}	skew	$\sigma_{\bar{x}}$	med.	range	α_{95}	Ch?
rand100	100	0.02	0.95	0.19	-0.1	0.1	0.11	-2.4/2.2	-0.28/0.39	N
rand50a	50	0.05	0.98	0.28	-0.4	0.14	0.16	-2.4/1.9	-0.28/0.39	N
rand50b	50	-0.01	0.94	0.27	0.1	0.13	0.03	-1.9/2.2	-0.4/0.4	N
rand20a	20	0.03	0.9	0.42	-1	0.2	0.19	-2.4/1.3	-0.11/0.39	Y
rand20b	20	0.39	0.95	0.42	-0.1	0.21	0.53	-0.9/1.9	-0.55/1.23	N
pop	50	4.9	5.38	1.53	2.4	0.76	3.34	.4/29.3	2.3/4.8	Y
pop -1	49	4.41	4.11	1.18	1.5	0.59	3.27	.4/17.6	2.3/4.7	Y
ln(pop)	50	1.11	1.01	0.29	0	0.14	1.21	-0.8/3.4	0.8/1.6	N
taxes	50	1140	343	97	2	48	1056	553/2674	993/1161	Y
tax -1	49	1109	265	76	0.8	38	1055	553/1993	993/1141	Y
deficit	30	10.6	7.5	2.8	0.2	1.4	11.3	-1.8/25.7	5.5/14.1	N
HS grad	50	75.1	7.4	2.1	-0.1	1.1	76.2	58/90	72.9/78.0	N
smoked	10	69.6	31.2	22.3	0.5	9.8	69.2	65.7/75.3	66.4/73.6	N
Anch T	12	35.2	16.8	10.7	0	4.9	35	13/58	18/54	N
bat avg	90	352	21	4.5	0.9	2.3	350	313/424	342/354	Y
bat -30	60	347	15	3.9	0.2	1.9	346	313/385	341/353	N

Example 1: random normal numbers of Figures 1 and 2.

The data of Figures 1 and 2 are drawn from a table of random normal numbers and therefore are about as close as one can get to perfectly random, normally distributed data. The true population mean is zero, and the true population standard deviation is one; data units therefore could be called 'true standard deviations'. We will consider five datasets: one with N=100 (Rand100), two with N=50 (Rand50a & Rand50b), and two with N=20 (Rand20a & Rand20b). Measurements within each dataset are independent of each other, but datasets are not strictly independent: the N=100 example is the combination of the two N=50 examples, and the two N=20 examples are included in the first N=50 example.

All five examples have a mean (Table 6) that is very close to the true population mean of zero; the largest departure is 0.4 units. As we might expect, the calculated 95% confidence limits for the true mean (α_{95}) include zero for all five examples. The α_{95} for Rand20b, however, barely includes the true mean of zero. At first it seems surprising that we have almost disproved something that we know to be true: that the true mean is zero. We should remember, however, that if we did this test on 20 datasets instead of 5, we would expect an average of one test to 'fail' at the 95% confidence level.

The histograms of Figure 2 show considerable apparent character change when compared either to each other or to a theoretical normal distribution. This variability is typical sampling variability for small samples. This visual variability is mirrored by a variability in calculated skewness: one of the five (Rand20a) actually fails the rule of thumb that skewness should be less than ±0.5 for normally distributed data. In spite of the apparently substantial departures from a simple normal distribution in the histograms, the standard deviation is fairly robust: the standard deviation of each is about the same (0.90-0.98) and close to the true population value of 1.0. By coincidence, all five standard deviations are less than the true value of 1.0; such a coincidence would be highly unlikely (1 chance in 2^5) if the five datasets were truly independent rather than subsets of each other. The interquartile range, which is less efficient than the standard deviation, is similar (1.32-1.47) for the three larger datasets but highly variable (0.62-1.86) for the 20-point samples.

Rand20a, the apparently skewed dataset, is also the only dataset for which Chauvenet's criterion allows us to reject a measurement as anomalous. This same measurement of -2.41 was in Rand100 and Rand50a, but it was not considered rejectable by application of Chauvenet's criterion to those two datasets because more extreme values are expected when N is larger. Obviously (in hindsight), even exceedingly scarce extreme values will occasionally show up in small samples, seeming more anomalous in the small sample than in a large sample. Chauvenet's criterion was incorrect in suggesting that the measurement be rejected from Rand20a.

In all five examples, the median lies farther from the true mean of zero than the arithmetic mean does. Thus for these samples from a normally distributed parent population, the median is a less efficient and therefore less accurate estimate of the true population average than is the mean. Similarly, the range varies substantially among the different examples, though we have seen that the standard deviation is relatively constant. For each of the five examples, the 95% confidence limits for the median are broader and therefore less efficient that 95% confidence limits for the mean; in every case these confidence limits for the median correctly includes the true population average of zero. Whether we use confidence limits for the mean or for the median, we see in Table 6 that making 100 measurements rather than 20 lets us narrow our uncertainties in estimating the true population average by 50% or more.

Example 2: race between the hare and the tortoise.

In an update of the ancient race between the hare and the tortoise, the tortoise won the race and yet the hare got a speeding ticket. Since the tortoise won, its 'average' speed must have been faster than the hare's more erratic pace. Use of a mean and standard deviation would be quite inappropriate for the hare. The hare had a bimodal speed of either zero (resting) or -- rarely -- extremely fast; probably the mean would be closer to the dominant zero peak and the standard deviation would imply some negative speeds. Sampling the hare's speed at uniform time intervals would give a completely different picture than if its speed were sampled at uniform distance intervals: according to the former it was usually resting, but according to the latter it was usually breaking the speed limit.

Example 3: percentage of high school students that graduate, by state.

We cannot expect values of any variable for different states of the U.S.A. to be truly independent: adjacent states or states with similar industries could be expected to give more similar values than distant states with different economic bases. We will proceed anyway, because such examples are illustrative and because it is fruitless to respond to a question like "What is the average percentage of students that graduate from U.S. high schools?" with the answer "It is impossible to say, because it is invalid to average such data."

Figure 5 shows that the distribution of high-school graduation rates appears to be approximately normal. Indeed, it looks more like a bell-shaped or Gaussian distribution than do Figures 1 and 2 (which are known to come from a normally distributed parent population). Furthermore, skewness is low, and Chauvenet's criterion does not reject any data. Thus it is relatively safe to conclude that the calculated average graduation percentage is 75.1% and that the 'true' average is 75.1±2.1%. Non-parametric statistics are neither needed nor as appropriate as parametric statistics for this dataset. The mean value of 75.1 is close to the median of 76.2, at least in comparison to the high standard deviation of 7.4, again suggesting normality.

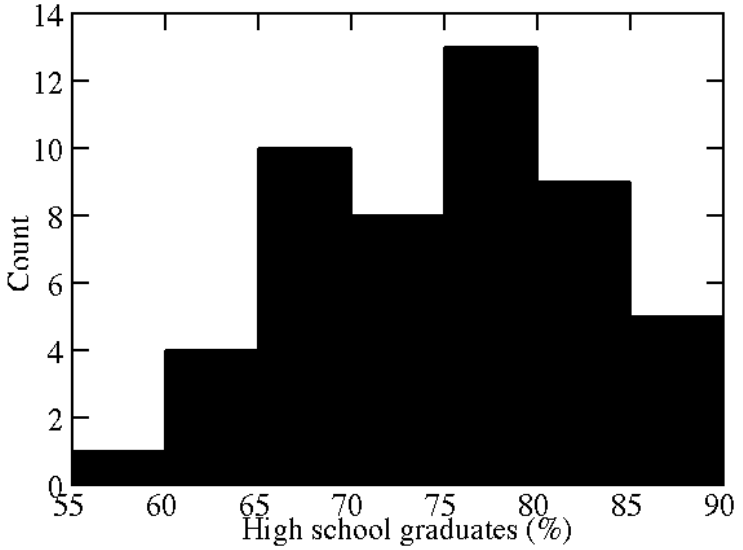


Figure 5. Percentage of high-school students who graduate, for U.S. states.

Example 4: population of U.S. states (1990 census).

The populations, in millions, of the U.S. states obviously diverge from a normal distribution (Figure 6a). Our 'quick-and-dirty' technique of comparing mean to median indicates a non-normal distribution: the mean is almost 50% larger than the median, and examination of Figure 6a suggests that one anomalously high value is at least partially responsible for pulling the mean so far to the right of the median. The distribution has a strong positive skewness of 2.4, with no left tail and a long right tail.

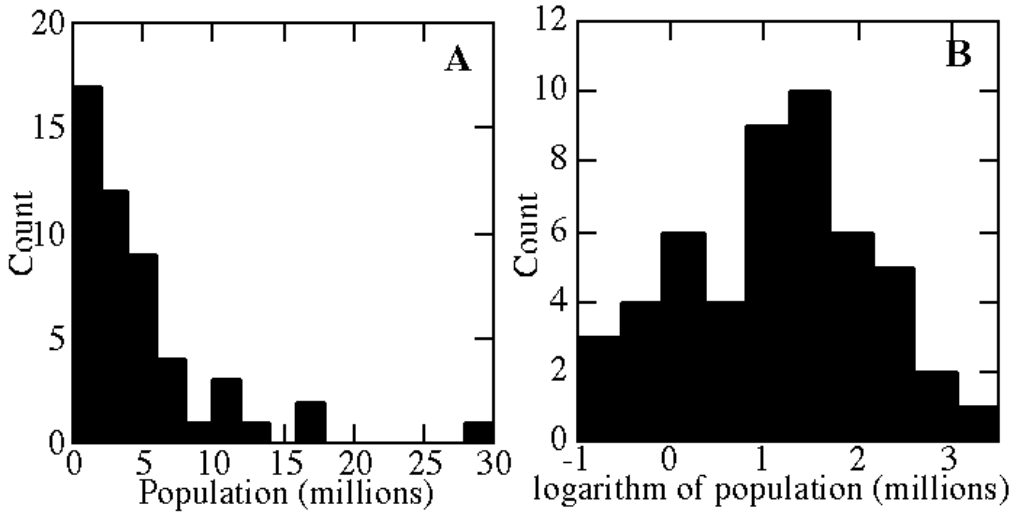


Figure 6. Populations of the U.S. states in 1990. Note that the highly skewed distribution of A is transformed to a nearly normal distribution by converting to logarithm of population (B).

Should the one extremely large value of 29,279,000 (29.3 million or 29.3M) for California population be excluded as anomalous? Chauvenet's criterion says that any value of >18.7M can be excluded, so 29.3M is far beyond the minimum cutoff. If we exclude California and recalculate mean and standard deviation, reapplication of Chauvenet's criterion (not recommended) would suggest that we reject two more states with large populations. I have not done so, though it might be interesting to see how many states we would ultimately exclude through repeated use of Chauvenet's criterion.

If one is statistically justified in excluding at least California, then such an exclusion implies that California is in some way unique or anomalous, with some different variable controlling its population than is operant (or at least important) for populations of the other states. As a former member of the California population, I can think of many ways in which one would describe the California population as anomalous, but that question is beyond the scopes of these data and of our concern. The key point is that the analysis flags an anomaly; it cannot explain the anomaly.

Figure 4 suggests that one's first reaction to a non-normal distribution should not be to discard data; it is to consider transforms that might convert the dataset to an approximately normal distribution. The most common transform is to take natural logarithms of the data, and the logarithmic transform is most likely to succeed in cases such as the present one that have a strong positive skewness. Figure 6b is such a transform. Logarithm of population visually does appear to be normally distributed, mean and median are similar (with a difference that is only about 10% of the standard deviation), and skewness is zero (!). Thus we may conclude that state population is log-normally distributed, with a mean of 3.0M ($e^{1.1}$, because the mean of the natural logarithms of population is 1.1).

Knowing that it is much more appropriate to analyze logarithms of state populations than raw state populations, we can now apply Chauvenet's test and find that no data should be excluded. Our previous temptation to exclude California was ill founded. With any logarithmic distribution the largest values tend to be more widely spaced than the smallest values. I suspect that Chauvenet's criterion will recommend exclusion of one or even many valid data points whenever a dataset has

strong skewness (positive or negative), because Chauvenet's criterion is excessively sensitive to such violations of normality.

Example 5: average state taxes per capita, by state.

What is the average amount of state taxes paid in the U.S.? One answer might come from simply dividing the total of all state tax income by the total U.S. population. Here is another, more informative approach.

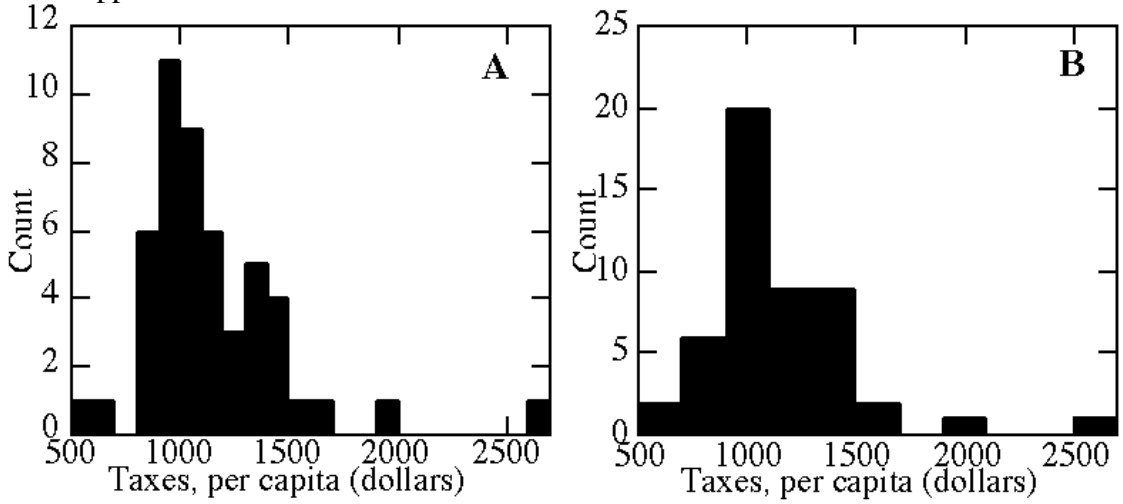


Figure 7. Histograms of per capita state taxes, for all U.S. states. Data are the same in A and B, but binning interval is much coarser in B.

Histograms of state taxes per capita, by state, are shown in Figures 7a and 7b. Although the two histograms show the same data, they emphasize slightly different features because of their different binning intervals. The coarser binning interval of Figure 7b makes the distribution look more normal, as is often the case for coarser binning (within reason). Finer binning (Figure 7a) makes the largest datum, per capita taxes of $2674 ($2.7K) in Alaska, look more anomalous. Both histograms are roughly bell-shaped but positively skewed (skewness=2.0). Thus it is worth trying a transform to logarithm of per capita taxes, but such a distribution is not shown because it accomplishes no improvement.

Chauvenet's criterion shows that the value for Alaska taxes is far more than is likely by chance for a normal distribution. Recalculation of parametric statistics after omitting Alaska gives a more normal value of skewness, and permits us to exclude yet another state (Hawaii), but we will forgo that opportunity. It seems likely that Alaska and possible that Hawaii are anomalous in state taxes in comparison to other states, because costs are greater at remote locations.

Chapter 3: Induction & Pattern Recognition

Induction is pattern recognition -- an inference based on limited observational or experimental data -- and pattern recognition is an addictively exhilarating acquired skill.

Of the two types of scientific inference, induction is far more pervasive and useful than deduction (Chapter 4). Induction usually infers some pattern among a set of observations and then attributes that pattern to an entire population. Almost all hypothesis formation is based consciously or subconsciously on induction.

Induction is pervasive because people seek order insatiably, yet they lack the opportunity of basing that search on observation of the entire population. Instead they make a few observations and generalize.

[Harris, 1982]

Induction is not just a description of observations; it is always a leap beyond the data -- a leap based on circumstantial evidence. The leap may be an inference that other observations would exhibit the same phenomena already seen in the study sample, or it may be some type of explanation or conceptual understanding of the observations; often it is both. Because induction is always a leap beyond the data, it can never be proved. If further observations are consistent with the induction, then they **confirm**, or lend substantiating support to, the induction. But the possibility always remains that as-yet-unexamined data might disprove the induction.

In symbols, we can think of confirmation of our inductive hypothesis A as: $A \Rightarrow B$, B, $\therefore A$ (i.e., A implies B; B is observed and therefore A must also be true or present). Such evidence may be inductively useful confirmation. The logic, however, is a deductive fallacy (known as affirming the consequent), because there may always be other factors that cause B. Although confirmation of an induction is incremental and inconclusive, the hypothesis can be disproved by a single experiment, via the deductive technique of modus tollens: $A \Rightarrow B$, $-B$, $\therefore -A$ (i.e., A implies B; B is not observed and therefore A must not be true or present).

Scientific induction requires that we make two unprovable assumptions, or postulates:

- **representative sampling**. Only if our samples are representative, or similar in behavior to the population as a whole, may we generalize from observations of these samples to the likely behavior of the entire population. In contrast, if our samples represent only a distinctive subset of the population, then our inductions cannot extend beyond this subset. This postulate is crucial, it is usually achieved easily by the scientist, and yet it is often violated with scientifically catastrophic results. As discussed more fully in the previous chapter, randomization and objective sampling are the paths to obtaining a representative sample; subjective sampling generates a biased sample.

- **uniformity of nature**. Strictly speaking, even if our sample is representative we cannot be certain that the unsampled remainder of the population exhibits the same behavior. However, we as-

sume that nature is uniform, that the unsampled remainder is similar in behavior to our samples, that today's natural laws will still be valid tomorrow. Without this assumption, all is chaos.

* * *

Types of Explanation

Induction is explanation, and explanation is identification of some type of order in the universe. Explanation is an integral part of the goal of science: perceiving a connection among events, deciphering the explanation for that connection, and using these inductions for prediction of other events. Some scientists claim that science cannot explain; it can only describe. That claim only pertains, however, to Aristotelian explanation: answering the question "Why?" by identifying the purpose of a phenomenon. More often, the scientific question is "How?" Here we use the inclusive concept of explanation as any identification of order.

Individual events are complex, but explanation discerns their underlying simplicity of relationships. In this section we will consider briefly two types of scientific explanation: comparison (analogy and symmetry) and classification. In subsequent sections we will examine, in much more detail, two more powerful types of explanation: correlation and causality.

Explanation can deal with attributes or with variables. An attribute is binary: either present or absent. Explanation of attributes often involves consideration of associations of the attribute with certain phenomena or circumstances. A variable, in contrast, is not merely present or absent; it is a characteristic whose changes can be quantitatively measured. Explanations of a variable often involve description of a correlation between changes in that variable and changes in another variable. If a subjective attribute, such as tall or short, can be transformed into a variable, such as height, explanatory value increases.

The different kinds of explanation contrast in explanatory power and experimental ease. Easiest to test is the **null hypothesis** that two variables are completely unrelated. Statistical rejection of the null hypothesis can demonstrate the likelihood that a classification or correlation has predictive value. Causality goes deeper, establishing the origin of that predictive ability, but demonstration of causality can be very challenging. Beyond causality, the underlying quantitative theoretical mechanism sometimes can be discerned.

* * *

Comparison is the most common means of identifying order, whether by scientists or by lay people. Often, comparison goes no farther than a consideration of the same characteristic in two individuals. Scientific comparison, however, is usually meant as a generalization of the behavior of variables or attributes. Two common types of comparison are symmetry and analogy.

Symmetry is a regularity of shape or arrangement of parts within a whole -- for example, a correspondence of part and counterpart. In many branches of science, recognition of symmetry is a useful form of pattern recognition. To the physicist, symmetry is both a predictive tool and a standard by which theories are judged.

> In his book on symmetry, physicist Hermann Weyl [1952] said: "Symmetry, as wide or as narrow as you may define its meaning, is one idea by which man through the ages has tried to comprehend and create order, beauty, and perfection."
>
> I've always been confident that the universe's expansion would be followed by a contraction. Symmetry demands it: big bang, expanding universe, gravitational decel-

> eration, contracting universe, big crunch, big bang, . . . No problem of what happens before the big bang or after the big crunch; an infinite cycle in both directions. The only concern was that not enough matter had been found to generate sufficient gravity to halt the expansion. But dark matter is elusive, and I was sure that it would be found. Now, however, this elegant model is apparently overthrown by evidence that the expansion is accelerating, not decelerating [Schwarzschild, 2001]. Symmetry and simplicity do not always triumph.

Analogy is the description of observed behavior in one class of phenomena and the inference that this description is somehow relevant to a different class of phenomena. Analogy does not necessarily imply that the two classes obey the same laws or function in exactly the same way. Analogy often is an *apparent* order or similarity that serves only as a visualization aid. That purpose is sufficient justification, and the analogy may inspire fruitful follow-up research. In other cases, analogy can reflect a more fundamental physical link between behaviors of the two classes. Either type of analogy can bridge tremendous differences in size or time scale. For example, the atom and the solar system are at two size extremes and yet their orbital geometries are analogous from the standpoints of both visualization and Newtonian physics. Fractals, in contrast, also describe similar physical phenomena of very different sizes, but they go beyond analogy by genetically linking different scales into a single class.

Analogy is never a final explanation; rather it is a potential stepping-stone to greater insight and hypothesis generation. Unfortunately, however, analogy sometimes is misused and treated like firm evidence. The following two examples illustrate the power of exact analogy and the fallacy of remote analogy.

> Annie Dillard [1974] on the analogy between chlorophyll and hemoglobin, the bases of plant and animal energy handling: "All the green in the planted world consists of these whole, rounded chloroplasts wending their ways in water. If you analyze a molecule of chlorophyll itself, what you get is one hundred thirty-six atoms of hydrogen, carbon, oxygen, and nitrogen arranged in an exact and complex relationship around a central ring. At the ring's center is a single atom of magnesium. Now: If you remove the atom of magnesium and in its exact place put an atom of iron, you get a molecule of hemoglobin."

> Astronomer Francesco Sizi's early 17th century refutation of Galileo's claim that he had discovered satellites of Jupiter [Holton and Roller, 1958]:
>
> "There are seven windows in the head, two nostrils, two ears, two eyes and a mouth; so in the heavens there are two favorable stars, two unpropitious, two luminaries, and Mercury alone undecided and indifferent. From which and many similar phenomena of nature such as the seven metals, etc., which it were tedious to enumerate, we gather that the number of planets is necessarily seven."

Comparison often leads to a more detailed explanation: **classification**. Classification can extract simple patterns from a mind-numbing quantity of individual observations, and it is also a foundation for most other types of scientific explanation. Classification is the identification of grounds for grouping complexly divergent individuals into a single class, based on commonality of some significant characteristic. Every individual is different, but we need and value tools for coping with this diversity by identifying classes of attributes. Indeed, many neurobiologists have concluded that people never experience directly the uniqueness of individual objects; instead, we unconsciously fit a suite of schemata, or classifications, to our perceptions of each object (Chapter 6).

A class is defined arbitrarily, by identifying a minimal number of characteristics required for inclusion in the class. Recognizing a scientifically useful classification, however, requires inductive insight. Ideally, only one or a few criteria specify a class, but members of the class also share many other attributes. For example, one accomplishes little by classifying dogs according to whether or not they have a scar on their ear. In contrast, classifying dogs as alive or dead (e.g., based on presence/absence of heartbeat) permits a wealth of generally successful predictions about individual dogs. Much insight can be gained by examining these ancillary characteristics. These aspects need not be universal among the class to be informative. It is sufficient that the classification, although based on different criteria, enhances our ability to predict occurrence of these typical features.

Classes are subjectively chosen, but they are defined according to objective criteria. If the criteria involve presence or absence of an attribute (e.g., use of chlorophyll), definition is usually straightforward. If the criteria involve a variable, however, the definition is more obviously subjective in its specification of position (or range of positions) along a continuum of potential values.

A classification scheme can be counterproductive [Oliver, 1991], if it imposes a perspective on the data that limits our perception. A useful classification can become counterproductive, when new data are shoved into it even though they don't fit.

Classifications evolve to regain utility, when exceptions and anomalous examples are found. Often these exceptions can be explained by a more restrictive and complex class definition. Frequently, the smaller class exhibits greater commonality of other characteristics than was observed within the larger class. For example, to some early astronomers all celestial objects were stars. Those who subdivided this class into 'wandering stars' (planets and comets) and 'fixed stars' would have been shocked at the immense variety that later generations would discover within these classes.

Each scientist applies personal standards in evaluating the scope and size of a classification. The 'splitters' favor subdivision into small subclasses, to achieve more accurate predictive ability. The 'lumpers' prefer generalizations that encompass a large portion of the population with reasonable but not perfect predictive accuracy. In every field of science, battles between lumpers and splitters are waged. For many years the splitters dominate a field, creating finer and finer classifications of every variant that is found. Then for a while the lumpers convince the community that the pendulum has swung too far and that much larger classes, though imperfect, are more worthwhile.

A class can even be useful though it has no members whatsoever. An **ideal class** exhibits behavior that is physically simple and therefore amenable to mathematical modeling. Even if actual individual objects fail to match exactly the defining characteristics of the ideal class, they may be similar enough for the mathematical relationships to apply. Wilson [1952] gives several familiar examples of an ideal class: physicists often model rigid bodies, frictionless surfaces, and incompressible fluids, and chemists employ the concepts of ideal gases, pure compounds, and adiabatic processes.

* * *

Coincidence

Classifications, like all explanations, seek meaningful associations and correlations. Sometimes, however, they are misled by coincidence.

> "A large number of incorrect conclusions are drawn because the possibility of chance occurrences is not fully considered. This usually arises through lack of proper controls and insufficient repetitions. There is the story of the research worker in nu-

> trition who had published a rather surprising conclusion concerning rats. A visitor asked him if he could see more of the evidence. The researcher replied, 'Sure, there's the rat.'" [Wilson, 1952]

Without attention to statistical evidence and confirmatory power, the scientist falls into the most common pitfall of non-scientists: **hasty generalization**. One or a few chance associations between two attributes or variables are mistakenly inferred to represent a causal relationship. Hasty generalization is responsible for many popular superstitions, but even scientists such as Aristotle were not immune to it. Hasty generalizations are often inspired by coincidence, the unexpected and improbable association between two or more events. After compiling and analyzing thousands of coincidences, Diaconis and Mostelle [1989] found that coincidences could be grouped into three classes:

- cases where there was an unnoticed causal relationship, so the association actually was not a coincidence;
- nonrepresentative samples, focusing on one association while ignoring or forgetting examples of non-matches;
- actual chance events that are much more likely than one might expect.

An example of this third type is that any group of 23 people has a 50% chance of at least two people having the same birthday.

Coincidence is important in science, because it initiates a search for causal relationships and may lead to discovery. An apparent coincidence is a perfectly valid source for hypotheses. Coincidence is not, however, a hypothesis test; quantitative tests must follow.

The statistical methods seek to indicate quantitatively which apparent connections between variables are real and which are coincidental. Uncertainty is implicit in most measurements and hypothesis tests, but consideration of probabilities allows us to make decisions that appropriately weigh the impact of the uncertainties. With suitable experimental design, statistical methods are able to deal effectively with very complex and poorly understood phenomena, extracting the most fundamental correlations.

* * *

Correlation

> "Every scientific problem is a search for the relationship between variables." [Thurstone, 1925]

Begin with two variables, which we will call X and Y, for which we have several measurements. By convention, X is called the independent variable and Y is the dependent variable. Perhaps X causes Y, so that the value of Y is truly dependent on the value of X. Such a condition would be convenient, but all we really require is the possibility that a knowledge of the value of the independent variable X *may* give us some ability to predict the value of Y.

* * *

To introduce some of the concerns implicit in correlation and pattern recognition, let's begin with three examples: National League batting averages, the government deficit, and temperature variations in Anchorage, AK.

Example 1: highest annual batting average in the National League.

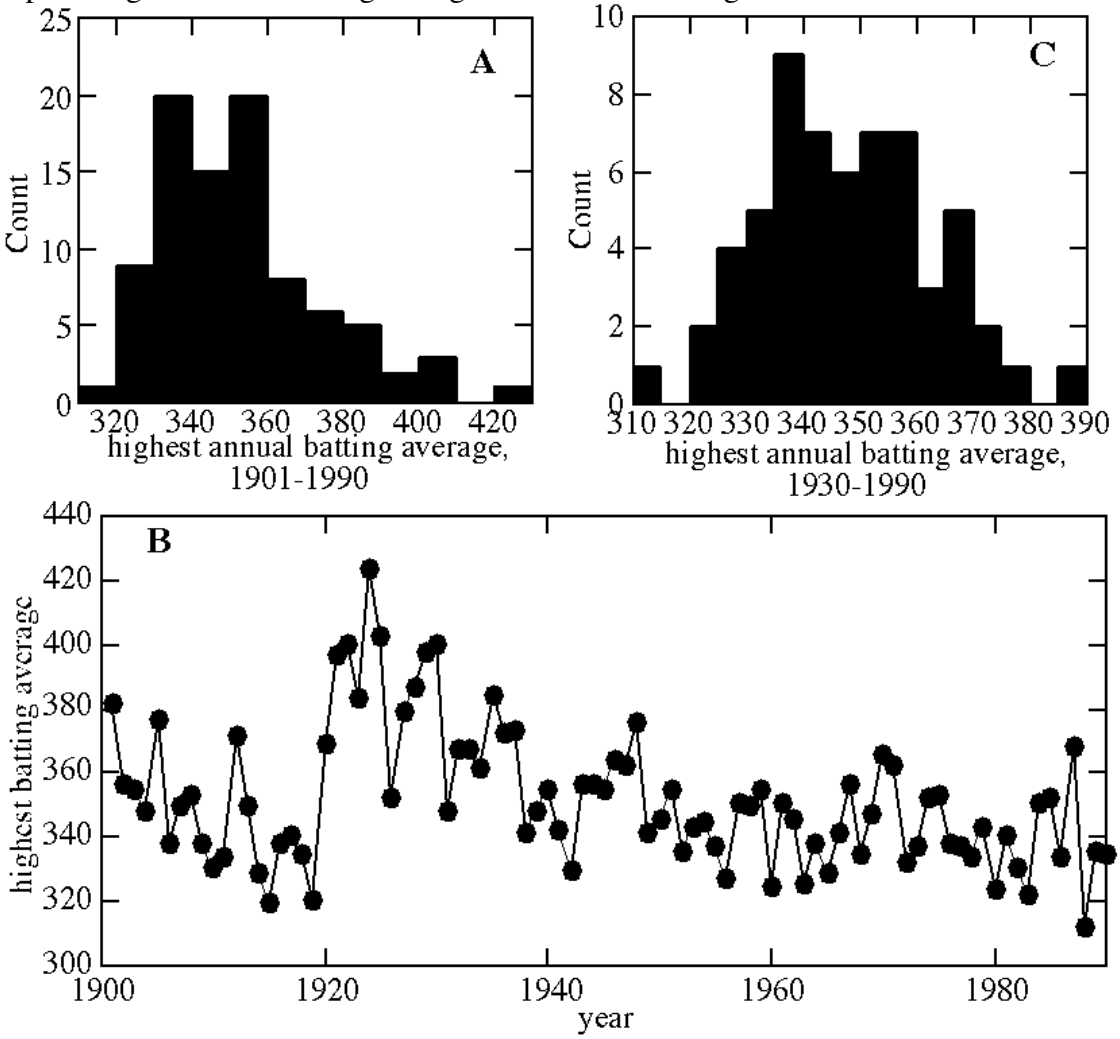


Figure 8. Highest annual batting average in the National League. Plotting results versus time (B) shows that the overall distribution for 1901-1990 (A) is skewed by periods of unusually low and high averages before 1930. Results for 1930-1990 (C) are more normally distributed.

We consider here the maximum batting average obtained by any National League player in each of the years 1901-1990. Because batting average is a time series, data certainly are not independent and we must beware of temporal trends. If we were to ignore the possibility of temporal trends, we would conclude that the data exhibit moderately normal behavior (Figure 8a), with slight positive skewness and, according to Chauvenet's criterion, one anomalously high value of 424 that could be excluded. Ignoring temporal trends, we would predict at a 68% confidence level (1σ) that the maximum 1991 batting average would be 352±21 (Table 6).

Plotting batting average versus time (Figure 8b), however, we see immediately that the departures from the mean were nonrandom. Batting averages decreased rapidly during 1901-1919, peaked during 1921-1930, and decreased gradually since then. What accounts for these long-term

trends? I am not enough of a baseball buff to know, but I note that the 1921-1930 peak is dominated by Rogers Hornsby, who had the highest average in 7 of these 10 years. Often in such analyses, identification of a trend's existence is the first step toward understanding it and, in some cases, toward preventing it.

Of course, substantial 'noise', or annual variation, is superimposed on these long-term trends. Later in this section, we will consider removal of such trends, but here we will take a simpler and less satisfactory approach: we will limit our data analysis to the time interval 1931-1990. We thereby omit the time intervals in which secular (temporal) trends were dominant. If this shorter interval still contains a slight long-term trend, that trend is probably too subtle to jeopardize our conclusions.

For 1931-1990 batting averages (Figure 8c), skewness is substantially less than for the larger dataset, and no points are flagged for rejection by Chauvenet's criterion. The standard deviation is reduced by one third, but the 95% confidence limits are only slightly reduced because the decrease in number of points counteracts the improvement in standard deviation.

Confining one's analysis to a subsample of the entire dataset is a legitimate procedure, *if* one has objective grounds for defining the subset and if one does not apply subset-based interpretations to the overall population. Obviously it would be invalid to analyze a 'subset' such as batting averages less than 400. Will the 1991 maximum batting average be 347±15 as predicted by the 1931-1990 data, or will there be another Rogers Hornsby?

Example 2: U.S. government deficit as a percentage of outlays, for 1960-1989.

Again we are dealing with a time series, so the flowchart of Figure 4 recommends that our first step is to plot deficit percentage versus time (Figure 9b). Such a plot exhibits a strong secular trend of increasing deficit percentage, on which is superposed more 'random' year-to-year variations. In other words, the major source of variance in deficits is the gradual trend of increasing deficit, and annual variations are a subsidiary effect. Because our data are equally spaced in time, the superposition of these two variances gives a blocky, boxcar-like appearance to the histogram (Figure 9a), with too little tail. If the secular trend were removed, residuals would exhibit a more bell-shaped distribution.

If we ignore the secular trend, nonparametric statistics are more appropriate for this dataset than are parametric statistics. However, ignoring the major source of variance in a dataset is almost always indefensible. Instead, a secular trend can be quantified and used to refine our understanding of a dataset. Later in this chapter, we will return to this example and determine that secular trend.

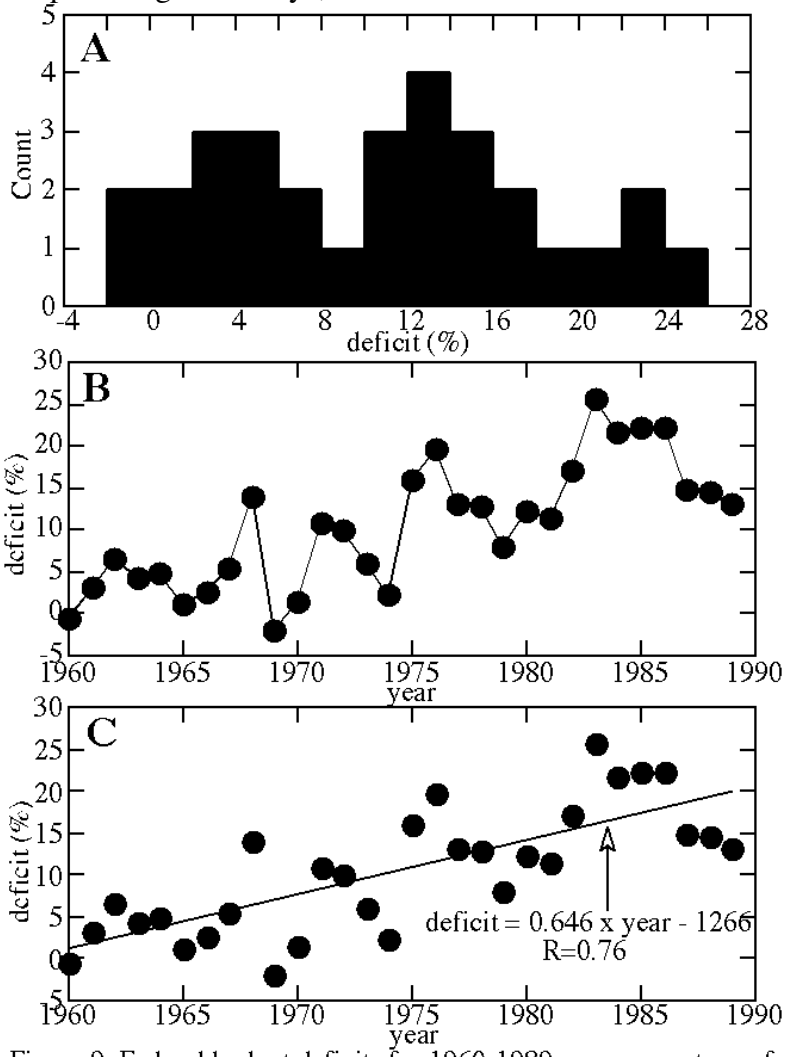


Figure 9. Federal budget deficits for 1960-1989, as a percentage of total budget.

Example 3: Monthly averages of temperature for Anchorage, Alaska.

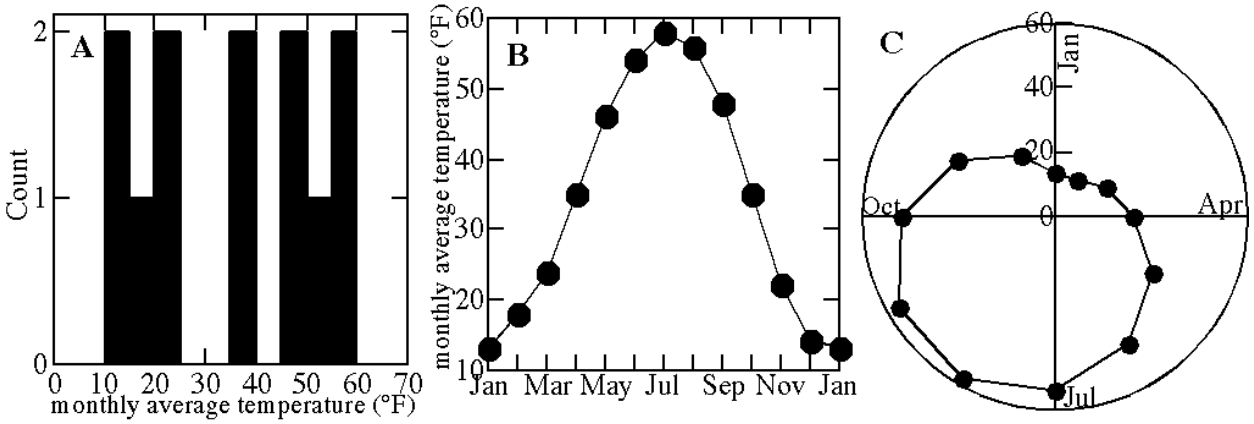


Figure 10. Monthly average temperatures (°F) for Anchorage, Alaska. A histogram display (A) of these data is useless. The pattern of temperature changes needs to be viewed versus time, in either a linear plot (B) or polar plot (C).

The histogram of monthly temperatures in Anchorage (Figure 10a) is strongly bimodal, with equal-sized peaks at 10-25° and at 45-60°. Skewness is zero because the two peaks are equal in size, so the mean is close to the median and both are a good estimate of the true average. Many bimodal distributions have one dominant peak, however, causing a distribution that is skewed and biasing both the mean and median.

Nonparametric statistics are much more appropriate here than parametric statistics. Neither is an acceptable substitute for investigation of the causes of a bimodal distribution. For this example, the answer lies in the temporal trends. Again we have a time series, so a plot of temperature versus time may lend insight into data variability. Months of a year can define an 'ordinal' scale: order along a continuum is known but there is neither a time zero nor implicitly fixed values. Here I simply assigned the numbers 1-13 to the months January-December-January for plotting, keeping in mind that the sequence wraps around so that January is both 1 and 13, then I replaced the number labels with month names (Figure 10b). A circular plot type known as polar coordinates is more appropriate because it incorporates wraparound (Figure 10c).

Consider the absurdities of simply applying parametric statistics to datasets like this one. We calculate that the average temperature is 35.2° (i.e., cold), but in fact the temperature almost never is cold. It switches rapidly from cool summer temperatures to bitterly cold winter temperatures. Considering just the standard deviation, we would say that temperature variation in Anchorage is like that in Grand Junction, Colorado (16.8° versus 18.7°). Considering just the mean temperature, we would say that the average temperature of Grand Junction (52.8°) is similar to that of San Francisco (56.8°). Thus temperatures in Grand Junction, Colorado are statistically similar to those of San Francisco and Anchorage!

* * *

Crossplots

Crossplots are the best way to look for a relationship between two variables. They involve minimal assumptions: just that one's measurements are reliable and paired (x_i, y_i). They permit use of an extremely efficient and robust tool for pattern recognition: the eye. Such pattern recognition and its associated brainstorming are a joy.

Crossplot interpretation, like any subjective pattern recognition, is subject to the 'Rorschach effect': the brain's bias toward 'seeing' patterns even in random data. The primary defense against the Rorschach effect is to subject each apparent pattern to some quantitative test, but this may be impractical. Another defense is to look at many patterns, of both random and systematic origins, in order to improve one's ability to distinguish between the two.

> Galison [1985] described the application of this approach in the bubble-chamber experiments at Berkeley. A computer program plotted histograms not only of the measured data but also of randomly generated pseudo-datasets. The investigator had to distinguish his datasets by recognizing which histograms had significant peaks. Louis Alvarez said that this program prevented many mistaken discovery claims and later retractions. Figure 2 makes me empathize with the problem faced by these particle physicists.

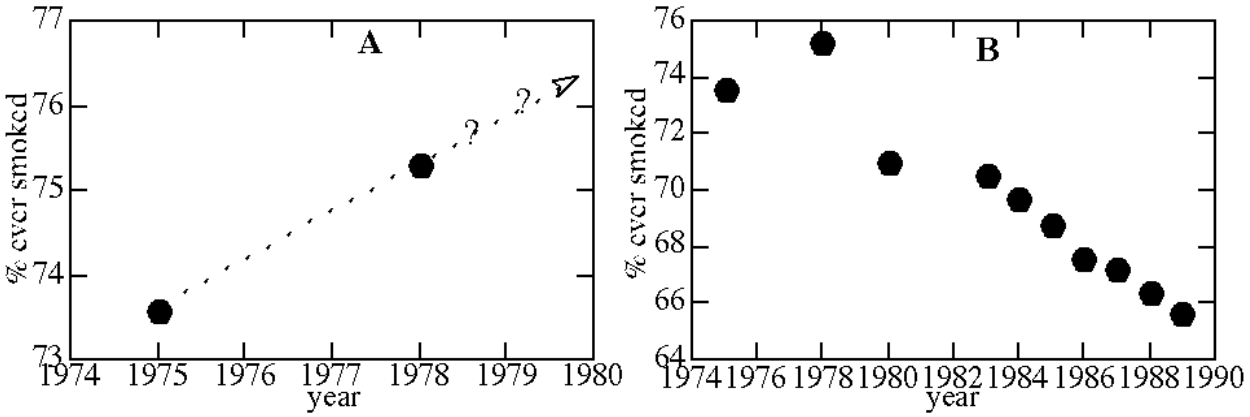


Figure 11. The hazards of extrapolation are shown by these plots of percentage of high school students who have smoked cigarettes. The apparent upward trend for 1975-1978 may be an artifact of less accurate data prior to 1982.

Data dispersion is inevitable with crossplots, and awareness of this dispersion is essential to crossplot interpretation. For example, consider the change through time of the percentage of American high school seniors who have ever smoked a cigarette. Figure 11a shows that this percentage increased from 73.6% to 75.3% in the three years from 1975 to 1978. If I were foolish enough to extrapolate from these two measurements, I could estimate that by the year 2022 100% of high school students will have tried cigarettes. The flaws are that one has no estimate of the errors implicit in these measurements and that extrapolation beyond the range of one's data is hazardous. As a rule of thumb, *it is moderately safe to extrapolate patterns to values of the independent variable that are perhaps 20% beyond that variable's measured range*, but extrapolation of Figure 11a to 2022 is more than an order of magnitude larger than the data range.

Figure 11b shows the eight subsequent determinations of percentage who have tried cigarettes. From this larger dataset it is evident that the apparent pattern of Figure 11a was misleading, and the actual trend is significantly downward. Based on these later results, we might speculate that one or both of the first two measurements had an error of about two percent, which masked a steady and possibly linear trend of decreasing usage. Alternatively, we might speculate that usage did increase temporarily. Is the steady trend of the rightmost seven points a result of improved polling techniques so that errors are decreased? Examination of such crossplots guides our considerations of errors and underlying patterns.

* * *

Plotting Hints

Crossplots can hide or reveal patterns. Plotting technique affects the efficiency of visual pattern recognition. Scientists are accustomed to a suite of plotting conventions, and they may be distracted if asked to look at plots that depart substantially from these conventions. I thank Open University [1970] for reminding me of some of the following plotting hints, which I normally take for granted. Figure 12 illustrates the effect of a few of these factors.

• Plot the dependent variable (the one whose behavior you hope to predict from the other variable) on the vertical axis, and plot the independent variable on the horizontal axis.

• Choose an aspect ratio for the plot that maximizes information (e.g., if we are examining the changes in *Y* values throughout a long time series, then the horizontal *X* axis can be much longer than the vertical *Y* axis).

• Plot variables with values increasing to the right and upward.

• Choose simple scale divisions, usually with annotated major divisions and with tics for simple subdivisions (e.g., range of 20-40 with annotation interval of 5 and tic spacing of 1).

• Choose a total plot range for each variable that is as small as possible, subject to these two restrictions: simple scale divisions and inclusion of all data points.

• Make an exception to the previous hint by including major meaningful scale divisions such as zero or 100%, only if this inclusion requires a relatively small expansion of the plot range.

• Plot data points as solid or open circles or crosses.

• If more than one dataset is included on the same plot, use readily distinguishable symbols.

• Label each axis with the variable name and its units.

• If data are a time series, connect the points with line segments. If they are independent, fit a line or curve

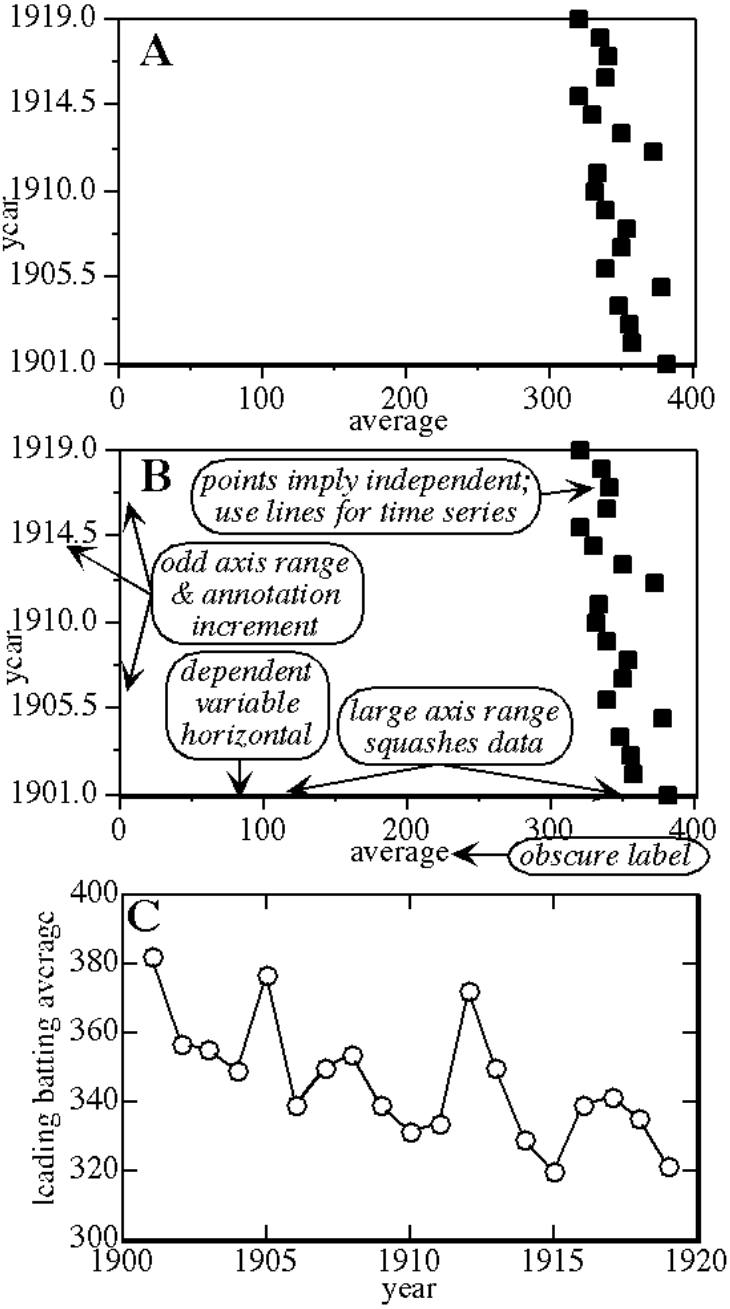


Figure 12. The same data are plotted in A and C, but poor choices of plotting parameters (B) in the top plot interfere with interpretation.

through the data, not connecting line segments.

• If possible, transform one or both variables so that the relationship between them is linear (e.g., choose among linear, semilog, and log-log plots).

Individual scientific specialties routinely violate one or more of the hints above. Each specialty also uses arbitrary unstated conventions for some plotting options:

• whether to frame the plot or just use one annotated line for each axis;

• whether to use an internal grid or just marginal tics on the frame or lines;

• whether to put tics on one or both sides, and whether to put them inside or outside the plot frame.

* * *

Extrapolation and Interpolation:

If a relationship has been established between variables X and Y, then one can predict the value of Y_i at a possibly unmeasured value of X_i. The reliability of this prediction depends dramatically on where the new X_i is with respect to the locations of the X_i that established the relationship. Several rules of thumb apply to interpolation and extrapolation:

• interpolation to an X_i location that is between closely spaced previous X_i is relatively safe,

• interpolation between widely spaced previous X_i is somewhat hazardous,

• extrapolation for a short distance (<20% of the range of the previous X_i) is somewhat hazardous,

• extrapolation for a great distance is foolhardy, and

• both interpolation and extrapolation are much more reliable when the relationship is based on independent data than when it is based on non-independent data such as a time series.

For example, when we saw the pattern of temporal changes in the U.S. deficit, the data appeared to fit a trend of increasing deficit rather well, so one should be able to extrapolate to 1991 fairly reliably. However, extrapolation ability is weaker for a time series than for independent events. As I am typing this, it is January 1991, the U.S. has just gone to war, Savings & Loans are dropping like flies, the U.S. is in a recession, and a deficit as small as the extrapolated value of 22% seems hopelessly optimistic. In contrast, when you read this, the U.S. budget hopefully is running a surplus.

As another example, we have already examined the changes with time of cigarette smoking among high school students, and we concluded that extrapolation from the two points of Figure 11a was foolhardy. With the data from Figure 11b, we might extrapolate beyond 1989 by perhaps 2-3 years and before 1975 by perhaps one year; the difference in confidence between these two extrapolations is due to the better-defined trend for 1983-1989 than for 1976-1980. Because these data are from a time series, any extrapolation is somewhat hazardous: if cigarette smoking were found in 1990 to be an aphrodisiac, the 1983-1989 pattern would immediately become an obsolete predictor of 1990 smoking rates. If there were such a thing as a class of 1986.5, then interpolation for the interval 1983-1989 would be very reliable (error <0.5%), because of extensive data coverage and small variance about the overall trend. In contrast, interpolation of a predicted value for some of the unsampled years in the interval 1975-1980 would have an error of at least 1%, partly because data spacing is larger but primarily because we are unsure how much of the apparent secular change

is due to measurement errors. If we knew that the first three measurements (in 1975, 1978, & 1980) constituted random scatter about the same well-defined trend of 1983-1989, then surprisingly it would be more accurate to predict values for these three years from the trend than to use the actual measurements.

An extreme example of the difference between extrapolation and interpolation for time series is world population (Figure 13). The validity of interpolated population within the last 2000 years depends on how much one trusts the simple pattern of Figure 13. The prolonged gap between 1 A.D. and 1650 conceivably could mask excursions as large as that of 1650-present, yet we know independently from history that such swings have not occurred. The combination of qualitative historical knowledge and the pattern of Figure 13 suggests that even the Black Death, which killed a large proportion of the population, caused less total change than is now occurring per decade. For purposes of defining the trend and for interpolation, then, both the distance between bracketing data points and the rate of change are important. Thus the great increase in sampling density at the right margin of Figure 13 is entirely appropriate, although a single datum at about 1000 A.D. would have lent considerable improvement to trend definition.

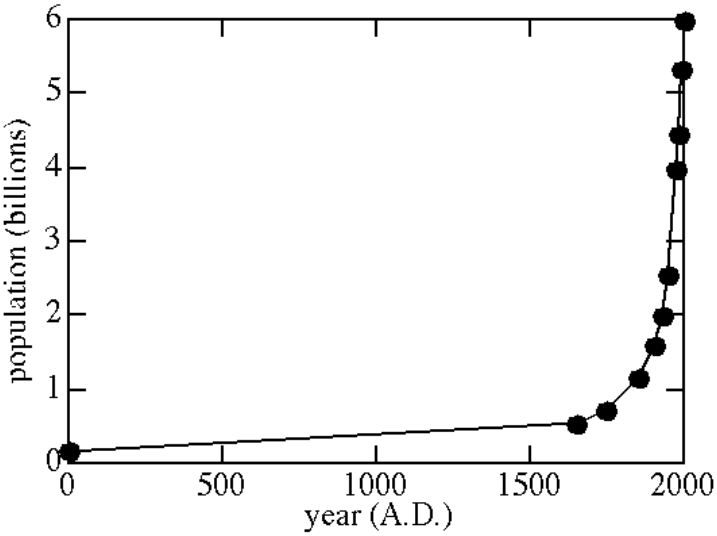


Figure 13. Growth in world population during the last 2000 years.

Extrapolation of world population beyond the limits of Figure 13 is both instructive and a matter of world concern. Predicting populations prior to 1 A.D. would be based on very scanty data, yet it appears that values would have been greater than zero and less than the 1 A.D. value of 0.2 billion. In contrast, extrapolation of the pattern to future populations suggests that the world population soon will be infinite. Reality intervenes to tell us that it is impossible for the pattern of Figure 13 to continue for much longer.

The three examples above are atypical in that they all are time series -- measurements of temporal changes of a variable. Interpolation, extrapolation, and indeed any interpretation of a time series is ambiguous, because time is an acausal variable. Often one can hypothesize a relationship between two variables that lends confidence to one's interpretation. In contrast, the source of variations within a time series may be unmeasured and possibly even unidentified.

The challenge of avoiding the confounding effect of time is present in all sciences. It is particularly acute within the social sciences, because some variables that might affect human behavior are difficult to hold constant throughout an experiment. For example, consider the relationship between height and weight of boys, shown in Figure 14a. The relationship is nonlinear, and we might be tempted to extrapolate that a 180-cm-high boy could be as much as twice as heavy as a 160-cm-high boy. Clearly neither height nor weight is normally distributed, and in fact it would be absurd to speak of the average height or weight of boys, unless one specified the boys' age. Figure 14a is actually based on a tabulation for boys of different ages. Age is the causal variable that controls both height and weight and leads to a correlation between the two. Both change systematically but

nonlinearly with age (Figures 14b and 14c): early growth is dominantly in height and later growth is dominantly in weight, leading indirectly to the pattern of Figure 14a.

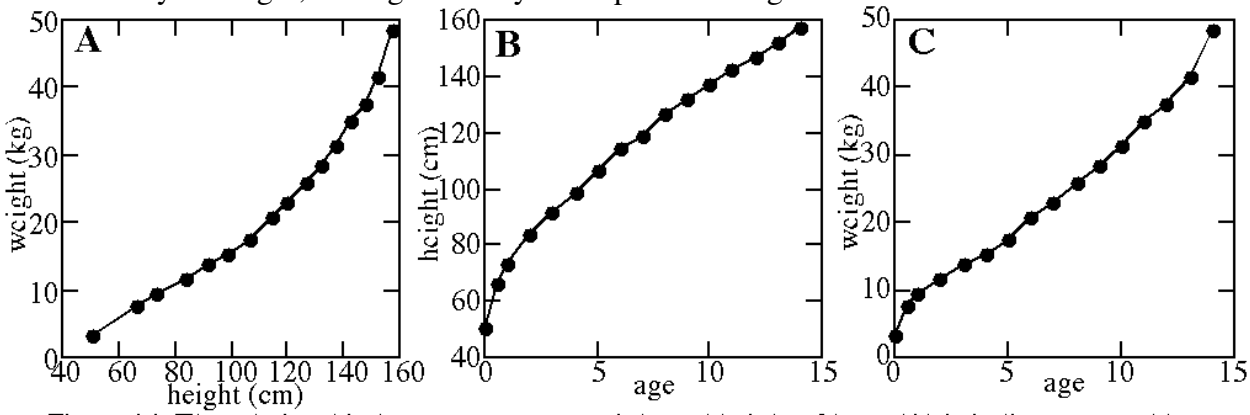


Figure 14. The relationship between average weight and height of boys (A) is indirect, caused by dependence of both on age (B & C).

Time series in particular, and nonindependent sampling in general, jeopardize interpolation and especially extrapolation. Nonlinearities are also a hazard, and we shall explore their impacts more fully in the subsequent section. First, however, let us assume the ideal correlation situation -- independent sampling and a linear relationship. How can we confidently and quantitatively describe the correlation between two variables?

* * *

Correlation Statistics

The type of test appropriate for identifying significant correlations depends on the kind of measurement scale. For classification data, such as male and female responses to an economic or psychological study, a test known as the contingency coefficient searches for deviations of observed from expected frequencies. For ranked, or ordinal, data where relative position along a continuum is known, the rank correlation coefficient is appropriate. Most scientific measurement scales include not just relative position but also measurable distance along the scale, and such data can be analyzed with the correlation coefficient or rank correlation coefficient. This section focuses on analysis of these continuous-scale data, not of classification data.

Suppose that we suspect that variable Y is linearly related to variable X. We need not assume existence of a direct causal relationship between the two variables. We do need to make the three following assumptions: first, that errors are present only in the Y_i; second, that these errors in the Y_i are random and independent of the value of X_i; and third, that the relationship between X and Y (if present) is linear. Scientists routinely violate the first assumption without causing too many problems, but of course one cannot justify a blunder by claiming that others are just as guilty. The second assumption is rarely a problem and even more rarely recognized as such. The third assumption, that of a linear relationship, is *often* a problem; fortunately one can detect violations of this assumption and cope with them.

The hypothesized linear relationship between X_i and Y_i is of the form: $\mathbf{Y = mX+b}$, where m is the slope and b is the Y intercept (the value of Y when X equals zero). Given N pairs of measurements (X_i,Y_i) and the assumptions above, then the slope and intercept can be calculated by **linear regression**, from:

$m = [N\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)]/[N\Sigma X_i^2 - (\Sigma X_i)^2]$

$b = [(\Sigma Y_i)(\Sigma X_i^2) - (\Sigma X_i Y_i)(\Sigma X_i)]/[N\Sigma X_i^2 - (\Sigma X_i)^2]$

Most spreadsheet and graphics programs include a linear regression option. None, however, mentions the implicit assumptions discussed above.

Linear regression fits the line that minimizes the squares of the residuals of Y_i deviations from the line. This concept is illustrated in Figure 15a, which shows a linear regression of leading National League batting averages for the years 1901-1920. This concept of minimizing the squares of Y_i deviations is very important to remember as one uses linear regression, for it accounts for several characteristics of linear regression.

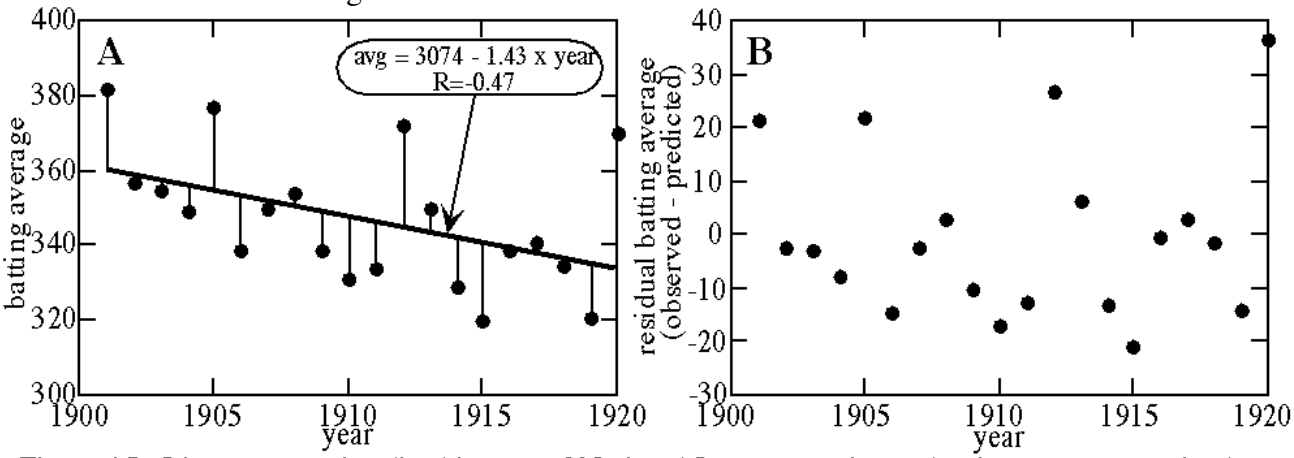


Figure 15. Linear regression (in this case of National League maximum batting average vs. time) minimizes the sum of squares of Y residuals (shown by vertical lines in A). Regression residuals (observed minus predicted values of Y), shown in B, are assumed to vary randomly about an average of zero and as a function of X (year).

First, we now understand the assumption that only the Y_i have errors and that these errors are random, for it is these errors or discrepancies from the trend that we are minimizing. If instead the errors were all in the X_i, then we should minimize the X_i instead (or, much easier, just rename variables so that Y becomes the one with the errors).

Second, minimizing the square of the deviation gives greatest weighting to extreme values, in the same way that extreme values dominate a standard deviation. Thus, the researcher needs to investigate the possibility that one or two extreme values are controlling the regression. One approach is to examine the regression line on the same plot as the data. Even better, plot the regression residuals -- the differences between individual Y_i and the predicted value of Y at each X_i, as represented by the vertical line segments in Figure 15a. Regression residuals can be plotted either as a function of X_i (Figure 15b) or as a histogram.

Third, the use of vertical deviations accounts for the name linear regression, rather than a name such as linear fit. If one were to fit a trend by eye through two correlated variables, the line would be steeper than that determined by regression. The best-fit line regresses from the true line toward a horizontal no-fit line with increases of the random errors of Y. This corollary is little-known but noteworthy; it predicts that if two labs do the same type of measurements of (X_i, Y_i), they will obtain different linear regression results if their measurement errors are different.

Fitting a linear regression does not imply that the obtained trend is significant. The **correlation coefficient (*R*)** measures the degree to which two variables are linearly correlated. We have seen above how to calculate the slope *m* of what is called the regression of *Y* on *X*: Y=mX+b. Conversely, we could calculate the slope *m'* of regression of *X* on *Y*: X=m'Y+b'. Note that we are abandoning the assumption that all of the errors must be in the Y_i. If X and Y are not correlated, then m=0 (a horizontal line) and *m'*=0 (a vertical line), so the product *mm'*=0. If the correlation is perfect, then *m*=1/*m'*, or *mm'*=1. Thus the product mm' provides a unitless measure of the strength of correlation between two variables [Young, 1962]. The correlation coefficient (*R*) is:

$$R=(mm')^{0.5} = [N\Sigma X_i Y_i-(\Sigma X_i)(\Sigma Y_i)]/\{[N\Sigma X_i^2-(\Sigma X_i)^2]^{0.5}\bullet[N\Sigma Y_i^2-(\Sigma Y_i)^2]^{0.5}\}$$

The correlation coefficient is always between -1 and 1. *R*=0 for no correlation, *R*=-1 for a perfect inverse correlation (i.e., increasing *X* decreases *Y*), and *R*=1 for a perfect positive correlation. What proportion of the total variance in *Y* is accounted for by the influence of *X*? R^2, a positive number between 0 and 1, gives that fraction.

Whether or not the value of R indicates a significant, or non-chance, correlation depends both on *R* and on *N*. Table 7 gives **95% and 99% confidence levels for significance of the correlation coefficient**. The test is called a two-tailed test, in that it indicates how unlikely it is that uncorrelated variables would yield either a positive or negative *R* whose absolute value is larger than the tabulated value. For example, linear regression of federal budget deficits versus time gives a high correlation coefficient of R=0.76 (Figure 9C). This pattern of steadily increasing federal budget deficits is significant at >99% confidence; for N=30, the correlation coefficient only needs to be 0.463 for the 99% significance level (Table 7).

Table 7: 95% and 99% confidence levels for significance of the correlation coefficient [Fisher and Yates, 1963].

N:	3	4	5	6	7	8	9	10	11	12
R_{95}:	0.997	0.95	0.878	0.811	0.754	0.707	0.666	0.632	0.602	0.576
R_{99}:	1	0.99	0.959	0.917	0.874	0.834	0.798	0.765	0.735	0.708
N:	13	14	15	16	17	18	20	22	24	26
R_{95}:	0.553	0.532	0.514	0.497	0.482	0.468	0.444	0.423	0.404	0.388
R_{99}:	0.684	0.661	0.641	0.623	0.606	0.59	0.561	0.537	0.515	0.496
N:	28	30	40	50	60	80	100	250	500	1000
R_{95}:	0.374	0.361	0.312	0.279	0.254	0.22	0.196	0.124	0.088	0.062
R_{99}:	0.479	0.463	0.402	0.361	0.33	0.286	0.256	0.163	0.115	0.081

Table 7 exhibits two features that are surprising. First, although we have already seen that *N*=2 gives us no basis for separating signal from noise, we would expect that *N*=3 or 4 should permit us to determine whether two variables are significantly correlated. Yet if *N*=3 or 4 we cannot be confident that the two variables are significantly correlated unless we find an almost perfectly linear correlation and thus an *R* of almost 1 or -1. Second, although we might accept that more pairs of (X_i, Y_i) points would permit detection of subtler correlations, it is still remarkable that with *N*>200 a cor-

relation can be significant even if *R* is only slightly larger than zero. With practice, one can tentatively identify whether two variables are significantly correlated by examining a crossplot, and Figure 16 is provided to aid that experience gathering. With very large *N*, however, the human eye is less able to identify correlations, and the significance test of Table 7 is much more reliable.

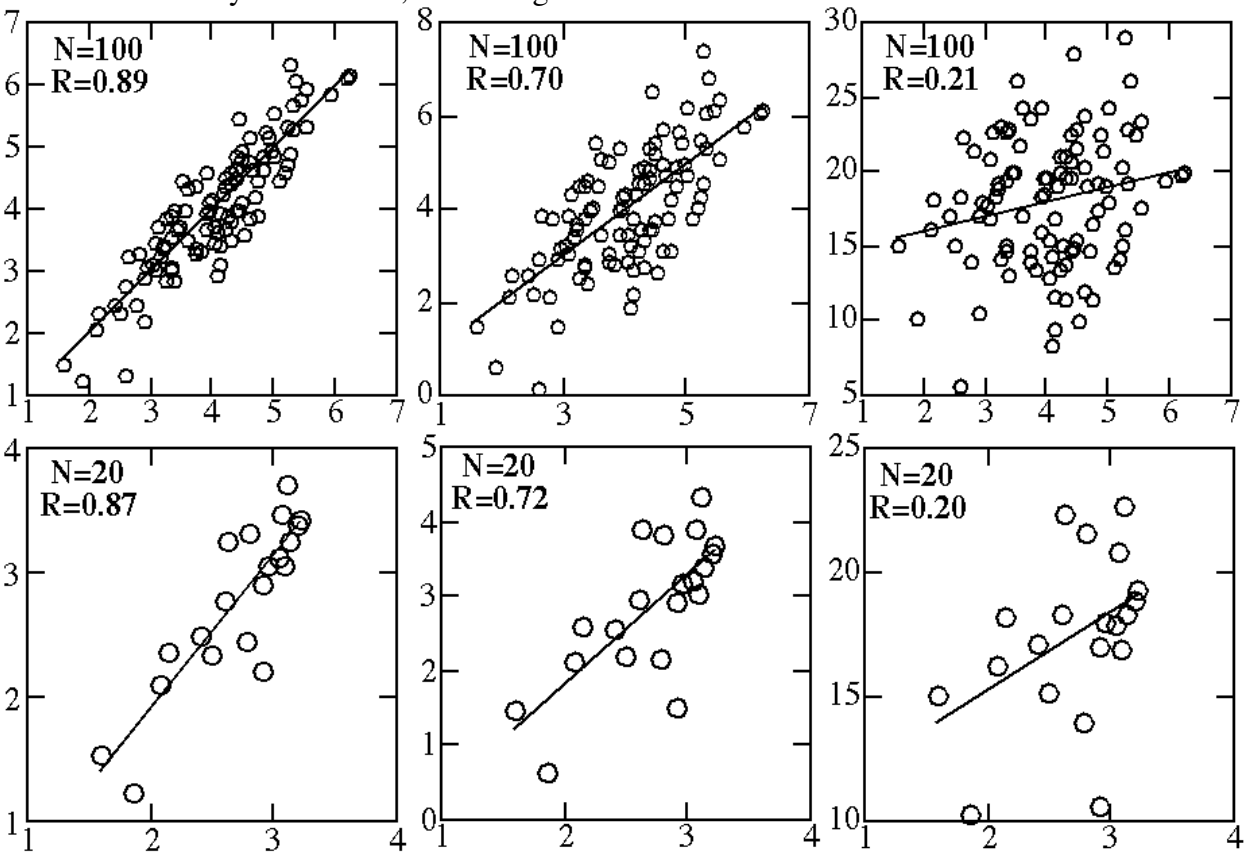


Figure 16. Examples of strong (R≈0.88), moderate (R≈0.71), and weak (R≈0.21) correlations, for N=100 points (top plots) and N=20 (bottom). Note that the linear regression line regresses toward horizontal as the correlation coefficient is reduced; only for the strongest correlation is this line as steep as would be drawn subjectively. All correlations except for that at lower right are significant at the 95% confidence level.

There is an adage: "One doesn't need statistics to determine whether or not two variables are correlated." This statement not only ignores scientists' preference for quantitative rather than qualitative conclusions; it is simply wrong when *N* is very small or very large. When *N* is very small (e.g., *N*<6), the eye sees correlations that are not real (significant). When *N* is very large (e.g., *N*>200), the eye fails to discern subtle correlations.

* * *

Nonlinear Relationships

The biggest pitfall of linear regression and correlation coefficients is that so many relationships between variables are nonlinear. As an extreme example, imagine applying these techniques to the annual temperature variation of Anchorage (Figure 10b). For a sinusoidal distribution such as this, the correlation coefficient would be virtually zero and regression would yield the absurd conclusion

that knowledge of what month it is (*X*) gives no information about expected temperature (*Y*). In general, any departure from a linear relationship degrades the correlation coefficient.

The first defense against nonlinear relationships is to transform one or both variables so that the relation between them is linear. Taking the logarithm of one or both is by far the most common transformation; taking reciprocals is another. Taking the logarithm of both variables is equivalent to fitting the relationship $Y=bX^m$ rather than the usual $Y=b+mX$. Our earlier plotting hint to try to obtain a linear relationship had two purposes. First, linear regression and correlation coefficients assume linearity. Second, linear trends are somewhat easier for the eye to discern.

A second approach is to use a nonparametric statistic called the **rank correlation coefficient**. This technique does not require a linear correlation. It does require a relationship in which increase in one variable is accompanied by increase or decrease in the other variable. Thus the technique is inappropriate for the Anchorage temperature variations of Figure 10b. It would work fine for the world population data of Figure 13, because population is always increasing but at a nonlinear rate. To determine the rank correlation coefficient, the steps are:

1) assign a rank to each X_i of from 1 to N, according to increasing size of X_i;

2) rank the Y_i in the same way;

3) subtract each X_i rank from its paired Y_i rank; we will call this difference in rankings d_i;

4) determine the rank correlation coefficient r, from

$$r = 1 - [6(\Sigma d_i^2)]/[N(N^2-1)].$$

The rank correlation coefficient r is much like the linear correlation coefficient R, in that both have values of -1 for perfect inverse correlation, 0 for no correlation, and +1 for perfect positive correlation. Furthermore, Table 7 above can be used to determine the significance of r in the same way as for R.

For example, the world population data of Figure 13 obviously show a close relationship of population to time. These data give a (linear) correlation coefficient of R=0.536, which is not significant according to Table 7. Two data transforms do yield correlations significant at the 99% confidence level: an exponential fit of the form $Y=b+10^{mx}$ (although this curve fit underestimates current population by more than 50%), and a polynomial fit of the form $Y=b+m_1X+m_2X^2$ (although it predicts that world population was much less than zero for 30-1680 A.D.!). In contrast, the rank correlation coefficient is r=1.0, which is significant at far more than the 99% confidence level.

Nonlinearities are common; the examples that we have just seen are a small subset. No statistical algorithm could cope with or even detect the profusion of nonlinear relationships. Thus I have emphasized the need to make crossplots and turn the problem of initial pattern recognition over to the eye.

Nonlinearities can be more sudden and less predictable than any of those shown within the previous examples. Everyone knows this phenomenon as 'the straw that broke the camel's back'; the scientific jargon is 'extreme sensitivity to initial conditions'. Chaos, a recent physics paradigm, now is finding such nonlinearities in a wide variety of scientific fields, particularly anywhere that turbulent motion occurs. The meteorologist originators of chaos refer to the 'Butterfly Effect': today's flapping of an Amazon butterfly's wings can affect future U.S. weather. Gleick [1987] gives a remarkably readable overview of chaos and its associated nonlinearities.

Due to extreme nonlinearities, a causal variable can induce a totally different kind of result at low concentration than at high concentration. An example is that nitroglycerin is a common medication for heart problems, yet the patient never explodes! Low concentrations of some causal variables can have surprisingly large effects, through development of a feedback cycle. Such a cycle, for example, is thought to account for the mechanism by which minute oscillations in the earth's orbit cause enormous fluctuations in global climate known as ice ages and interglacial stages. Extreme nonlinearities are the researcher's bane.

* * *

Correlation Conclusions

• Correlation can describe a relationship, but it cannot establish causality.

• Many variables have secular trends, but the correlation with time is indirect: secular change in a possibly unidentified causal variable causes the measured dependent variable to exhibit secular change.

• Crossplots are the most robust and reliable way to look for a relation between variables.

• Statistical correlation techniques assume independent measurements, so they must be used with caution when measurements are not independent (e.g., time series or grouped data).

• Interpolation between independent measurements is safe, but interpolation between non-independent measurements is risky.

• Extrapolation beyond the range of previous measurements is usually risky.

• Linear regression and the correlation coefficient *R* assume a linear relationship between variables.

• Examination of regression residuals is needed, to detect systematic mismatches.

• Nonlinearity can complicate relationships among variables enormously.

* * *

Perspectives on Causality

"Felix qui potuit rerum cognoscere causas."
(Happy is he who has been able to learn the causes of things) [Virgil, 70-19 B.C.]

Causality is a foundation of science, but it is not a firm foundation. Our concept of causality has been transformed more than once and it continues to evolve.

During the classical Greek period, to seek causes meant to seek the underlying purposes of phenomena. This concept of causality as purpose is identified with Aristotle, but Aristotle was an advocate rather than an initiator of this focus. The search for underlying purpose is also a religious concern, and the overlap between science and religion was correspondingly greater in ancient Greece than in modern times. Perhaps the religious connotation partly explains the shift away from Aristotelian causality during the last few centuries, but I suspect that the decisive factor was the growing scientific emphasis on verifiability. Greek science felt free to brainstorm and speculate about causes, but modern science demands tests of speculations. Testing purposes is much less feasible than testing modern associative causality. Modern scientific concern about purpose is confined primarily to some aspects of biology and social science. Even most of these questions (e.g.,

"what is the purpose of brain convolutions?") refer not to an underlying plan but to function or evolutionary advantage.

Hume [1935] redefined causality in more pragmatic terms. His definition of a cause is "an object precedent and contiguous to another, and where all objects resembling the former are placed in like relations of precedency and contiguity to those objects that resemble the latter." We can forgive Hume's constipated wording, I hope, on the grounds that definitions, like legal jargon, must be unambiguous and yet comprehensive. In other words, systematic (nonrandom) proximity in both space and time implies causality, and the event that occurs first is considered to be the cause of the second event. If event B is commonly preceded by an association with event A, then event A is a cause of event B. Note that neither requires the other: A may not be the only type of event that can cause B, and other conditions may be needed before A can cause B. We will consider these variables of causality and their tests in a later section on Mill's canons.

Lenzen [1938] used the example of Newtonian physics to demonstrate that even Hume's careful definition has exceptions. Cause does not always precede effect, as is evidenced by the fact that force causes simultaneous acceleration, not delayed acceleration. Cause and effect need not be contiguous, as is evidenced by the fact that gravitational attraction acts over millions of miles (else the earth would go careening away from the sun and off into space). To me, these exceptions are inconsequential. Hume's causality is meant to be a pragmatic concept, and a principle that is almost always useful should not be discarded for the purity of a void.

If causality is to be limited to the observable and testable as Hume's concept is, then several familiar attributes of causality may have to be stripped away: Aristotelian interest in purpose, the inevitability or necessity of an effect given a cause, and concern with underlying (unobservable) mechanisms [Boyd, 1985]. We are left with a sterile association between events, firmly founded in observations but lacking deeper understanding of processes. One twentieth-century philosophical school reached a similar conclusion with different logic: causality is nonunique -- one 'cause' can generate several paths and different causes can lead to the same 'effect' -- so causality should be confined to associations. Physicist Victor Weisskopf often said that causality is simply connections. A philosophical movement called logical positivism skirted this limitation by emphasizing that deduction from natural laws can provide a causal explanation of observations.

For the Sufis, cause-and-effect is a misguided focus on a single thread in the tapestry of intertwined relationships. They illustrate this lesson with the parable of the hanged man [Shah, 1972], which we can recast as follows:

> In 212 B.C., in his home in Syracuse, while working a math problem, Archimedes was killed by a Roman soldier. What caused his death? Was it that his applied scientific contributions – in the form of novel defensive weapons – were no defense against treason? Was it that the leader of the victorious invaders, in giving the order to leave the house of Archimedes alone, failed to assure that individual soldiers attended to the order? Was it that Archimedes, when commanded by a soldier to leave his home, was so preoccupied by his math problem that he refused to let even the fall of a city distract him? Or was it simply that the soldier had had a hard day, exhausting his patience for the cranky stubbornness of an old man?

* * *

Causality or pattern – is the choice a cultural one rather than innate? And if it is cultural, what about related fundamental scientific assumptions: comparison, linear thought, and time? A provocative perspective on these questions was provided by Lee's [1950] classic study of the language of the Trobriand Islanders, a virtually pristine stone-age culture of Southeast Asia. Her goals were to

extract both cultural information and fundamental insights into their thought patterns and "codification of reality". She did not assume that reality is relative; she did assume that different cultures can categorize or perceive reality in different ways, and that language provides clues to this perceptual approach.

The Trobriand language has no adjectives; each noun contains a suite of implicit attributes, and changing an attribute changes the object or noun. The Trobriand language has no change, no time, no distinction between past and present. Lacking these, it also lacks a basis for causality, and indeed there is no cause-and-effect. Absences of adjectives, change, and time distinctions are aspects of a broader characteristic: the virtual absence of comparisons of any kind in Trobriand language or world-view. There is no lineal connection between events or objects.

The Trobriand culture functions well without any of these traits that we normally consider essential and implicit to human perception. So implicit are these assumptions that Bronislaw Malinowski studied the Trobriand Islanders without detecting how fundamentally different their world-view is. Not surprisingly, he was sometimes frustrated and confused by their behavior.

The Trobriand people use a much simpler and more elegant perceptual basis than the diverse assumptions of change, time distinctions, causality, and comparison. *They perceive patterns*, composed of "a series of beings, but no becoming" or temporal connection. When considering a patterned whole, one needs no causal or temporal relationships; it is sufficient merely to identify ingredients in the pattern.

> "Trobriand activity is patterned activity. One act within this pattern brings into existence a pre-ordained cluster of acts. . . pattern is truth and value for them; in fact, acts and being derive value from the embedding pattern. . . To him value lies in sameness, in repeated pattern, in the incorporation of all time within the same point." [Lee, 1950]

During the last 12,000 years an ice age has waned, sea levels have risen, and climates have changed drastically. Plant and animal species have been forced to cope with these changes. Since the development of agriculture about 12,000 years ago, human progress has been incredibly fast. Biological evolution cannot account for such rapid human change; cultural evolution must be responsible. Perhaps the Trobriand example lends some insight into these changes. A Trobriandesque world-view, emphasizing adherence to pattern, might have substantial survival value in a stable environment. In contrast, climatic stress and changing food supplies favored a different world-view involving imagination and choice. Only in rare cases, such as the Trobriand tropical island, was the environment stable enough for a Trobriand-style perspective to persist.

The Trobriand world-view is in many ways antipodal to that upon which scientific research is based. Yet it is valid, in the same sense that our western world-view is valid: it works (at least in a stable environment). And the viability of such an alien perspective forces us to recognize that some of our fundamental scientific assumptions are cultural: our concepts of causality, comparison, and time may be inaccurate descriptions of reality.

* * *

Scientific causality transcends all of these restricted concepts of causality. It does not abandon concern with inevitability or with underlying mechanisms. Instead it accepts that description of causal associations is intrinsically valid, while seeking fundamental conceptual or physical principles that explain these associations.

Different sciences place different emphases on causality. The social sciences in general give a high priority to identifying causal relationships. Physical sciences often attempt to use causality as a

launching point for determining underlying theoretically-based quantitative relationships. Possibly this difference reflects the greater ease of quantifying and isolating variables in the physical sciences. Such sweeping generalizations are simplistic, however -- economics is an extremely quantitative social science.

All concepts of cause-and-effect assume that identical sets of initial conditions yield identical effects. Yet, quantum mechanics demonstrates that this fundamental scientific premise is invalid at the scale of individual atoms. For example, radioactive decay is intrinsically unpredictable for any one atom. If certainty is impossible at the atomic level, the same must be true for larger-scale phenomena involving many atoms. Werner Heisenberg, a champion of atomic-scale indeterminacy, carried this logic to a conclusion that sounds almost like a death knell for causality [Dillard, 1974]: "method and object can no longer be separated. The scientific world-view has ceased to be a scientific view in the true sense of the word."

Some non-scientists have seized on Heisenberg's arguments as evidence of the inherent limitations of science. Heisenberg's indeterminacy and the statistical nature of quantum mechanics are boundary conditions to causal description of particle physics, but not to causal explanation in general. Particle physicists emphasize that virtual certainty can still be obtained for larger-scale phenomena, because of the known statistical patterns among large numbers of random events. The pragmatic causality of scientists finds atomic indeterminacy to be among the least of its problems. Far more relevant is the overwhelming complexity of nature. Heisenberg may have shaken the foundations of science, but few scientists other than physicists felt tremors in the edifice.

It seems that a twentieth-century divergence is occurring, between theoretical concepts of causality and the working concepts used by scientists. One can summarize the differences among these different concepts of causality, using the following symbols:

A is the cause,
B is the effect,
$\Rightarrow$ means 'causes',
$\not\Rightarrow$ means 'does not necessarily cause',
$\therefore$ means 'therefore',
A_i is an individual observation of A, and
$\bar{A}$ is average behavior of A.

The different concepts of causality are then:

Sufi and Trobriand patterns: . . . . A, B, ...

Aristotle: $A \Rightarrow B$, in order to . . .

Hume: If A, then B; or A, $\therefore$B

logical positivist: theory C predicts 'A, $\therefore$B', & observation confirms it

Quantum mechanics: $A_i \not\Rightarrow B_i$, yet $\bar{A} \Rightarrow \bar{B}$

scientific consensus: If A, then probably B, possibly because . . . .

Scientists' working concept of causality remains unchanged, effectively useful, and moderately sloppy: *if one event frequently follows another, and no third variable is controlling both, then infer causality and, if feasible, seek the underlying physical mechanism*. Ambiguities in this working

concept sometimes lead to unnecessary scientific debates. For example, the proponent of a causal hypothesis may not expect it to apply universally, whereas a scientist who finds exceptions to the hypothesis may announce that it is disproved.

* * *

The logician's concept of causality avoids the ambiguity of the scientist's concept. Logicians distinguish three very different **types of causality: sufficient** condition, **necessary** condition, and a condition that is **both necessary and sufficient**.

If several factors are required for a given effect, then each is a *necessary* condition. For the example of Archimedes' death, both successful Roman invasion and his refusal to abandon his math problem were necessary conditions, or necessary causal factors. Many necessary conditions are so obvious that they are assumed implicitly. If only one factor is required for a given effect, then that factor is a *sufficient* condition. If only one factor is capable of producing a given effect, then that factor is a *necessary and sufficient* condition. Rarely is nature simple enough for a single necessary and sufficient cause; one example is that a force is a necessary and sufficient condition for acceleration of a mass.

Hurley [1985] succinctly describes the type of causality with which the scientist often deals:

> "Whenever an event occurs, at least *one* sufficient condition is present and *all* the necessary conditions are present. The conjunction of the necessary conditions *is* the sufficient condition that actually produces the event."

For the most satisfactory causal explanation of a phenomenon, we usually seek to identify the necessary and sufficient conditions, not a single necessary and sufficient condition. Often the researcher's task is to test a hypothesis that *N* attributes are needed (i.e., both necessary and sufficient) to cause an effect. The scientist then needs to design an experiment that demonstrates both the presence of the effect when the *N* attributes are present, and the absence of the effect whenever any of these attributes is removed.

Sometimes we cannot test a hypothesis of causality with such a straightforward approach, but the test is nevertheless possible using a logically equivalent statement of the problem. The following statements are logically equivalent [Hurley, 1985], regardless of whether *A* is the cause and *B* is the effect or vice versa (with -A meaning 'not-A' and ≡ meaning 'is equivalent to'):

A is a necessary condition for B
≡ B is a sufficient condition for A
≡ If B, then A (i.e., B, ∴A)
≡ If A is absent, then B is absent (i.e., -A, ∴-B)
≡ Absence of A is a sufficient condition for the absence of B
≡ Absence of B is a necessary condition for absence of A.

* * *

Mill's Canons: Five Inductive Methods

John Stuart Mill [1930], in his influential book System of Logic, systematized inductive techniques. The results, known as 'Mill's Canons', are five methods for examining variables in order to identify causal relationships. These techniques are extremely valuable and they are routinely used in modern scientific experiments. They are not, however, magic bullets that invariably hit the target.

The researcher needs to know the strengths and limitations of all five techniques, as each is most appropriate only in certain conditions.

A little jargon will aid in understanding the inductive methods. **Antecedent** conditions are those that 'go before' an experimental result; antecedent variables are those variables, known and unknown, that may affect the experimental result. **Consequent** conditions are those that 'follow with' an experimental result; consequent variables are those variables whose values are affected by the experiment. In these terms, the inductive problem is expressed as seeking the antecedent to the consequent of interest, i.e., seeking the causal antecedent. In considering the inductive methods, a useful shorthand is to refer to antecedent variables with the lower-case letters a, b, c, . . . and to refer to consequent variables with the upper-case letters Z, Y, X, . . .

Mill's Canons bear 19th-century names, but the concepts are familiar to ancient and modern people in less rigorous form:

a must cause *Z*, because:

- whenever I see *Z*, I also find *a* (*the method of agreement*);
- if I remove *a*, *Z* goes away (*the method of difference*);
- whether present or absent, *a* always accompanies *Z* (*the joint method of agreement and difference*);
- if I change *a*, *Z* changes correspondingly (*the method of concomitant variations*);
- if I remove the dominating effect of *b* on *Z*, the residual *Z* variations correlate with *a* (*the method of residues*).

Each of the five inductive methods has strengths and weaknesses, discussed below. The five methods also share certain limitations, which we will consider first.

Mill was aware that association or correlation does not imply causality, regardless of inductive method. For example, some other variable may cause both the antecedent and consequent ($h \Rightarrow c$, $h \Rightarrow Z$, $\therefore$ c correlates with Z, but $c \not\Rightarrow Z$). Thus Mill would expand the definition of each method below, ending each with an escape clause such as "or the antecedent and result are connected through some fact of causation." In contrast, I present Mill's Canons as methods of establishing relationships; whether the relationships are directly causal is an independent problem.

When we speak of a causal antecedent, we usually think of a single variable. Instead, the 'causal antecedent' may be a conjunction of two or more variables; we can refer to these variables as the primary and facilitating variables. If we are aware of the facilitating variables, if we assure that they are present throughout the experiment, and if we use the inductive methods to evaluate the influence of the primary variable, then success with Mill's Canons is likely. If we are unaware of the role of the facilitating variables, if we cannot turn them on and off at will, or if we cannot measure them, then we need a more sophisticated experimental design.

Method of Agreement

If several different experiments yield the same result, and these experiments have only one factor (antecedent) in common, then that factor is the cause of the observed result. Symbolically,

abc⇒Z, cde⇒Z, cfg⇒Z, ∴c⇒Z; or abc⇒ZYX, cde⇒ZW, cfg⇒ZVUT, ∴c⇒Z. The method of agreement is theoretically valid but pragmatically very weak, for two reasons:

• almost never can we be certain that the various experiments share only one common factor. We can increase confidence in the technique by making the experiments as different as possible (except of course for the common antecedent), thereby minimizing the risk of an unidentified common variable; and

• some effects can result from two independent causes, yet this method assumes that only one cause is operant. If two or more independent causes produce the same experimental result, the method of agreement will incorrectly attribute the cause to any antecedent that coincidentally is present in both experiments. Sometimes the effect must be defined more specifically and exclusively, so that different causes cannot produce the same effect.

It is usually safest to restate the method of agreement as: if several different experiments yield the same result, and these experiments *appear to* have only one antecedent factor in common, then that factor *may be* the cause of the observed result. Caution is needed, to assure that the antecedent and result are not both controlled by some third variable, that all relevant factors are included, and that the effect or result is truly of the same kind in all experiments. Time is a variable that often converts this method into a pitfall, by exerting hidden control on both antecedents and results. Ideally, the method of agreement is used only to spot a possible pattern, then a more powerful experimental design is employed to test the hypothesis.

Method of Difference

If a result is obtained when a certain factor is present but not when it is absent, then that factor is causal. Symbolically, abc⇒Z, ab⇒–Z, ∴c⇒Z; or abc⇒ZYXW, ab⇒YXW, ∴c⇒Z. The method of difference is scientifically superior to the method of agreement: it is much more feasible to make two experiments as similar as possible (except for one variable) than to make them as different as possible (except for one variable).

The method of difference has a crucial pitfall: no two experiments can ever be identical in all respects except for the one under investigation. Thus one risks attributing the effect to the wrong factor. Consequently, almost never is the method of difference viable with only two experiments; instead one should do many replicate measurements.

The method of difference is the basis of a powerful experimental technique: the controlled experiment. In a controlled experiment, one repeats an experiment many times, randomly including or excluding the possibly causal variable 'c'. Results are then separated into two groups -- experiment and control, or c-variable present and c-variable absent -- and statistically compared. A statistically significant difference between the two groups establishes that the variable *c* does affect the results, unless:

• the randomization was not truly random, permitting some other variable to exert an influence; or

• some other variable causes both *c* and the result.

> During his long imprisonment, the scientist made friends with a fly and trained it to land on his finger whenever he whistled. He decided to carry out a controlled experiment. Twenty times he whistled and held out his finger; every time the fly landed there. Then he pulled off the fly's wings. Twenty times he whistled and held out his finger; not once did the fly land there. He concluded that flies hear through their wings.

Joint Method of Agreement and Difference

If a group of situations has only one antecedent in common and all exhibit the same result, and if another group of similar situations lacks that antecedent and fails to exhibit the result, then that antecedent causes the result. Symbolically, abc⇒ZYX, ade⇒ZWV, and afg⇒ZUT; bdf⇒YWU and bceg⇒XVT, ∴a⇒Z.

This method is very similar to the methods of agreement and of difference, but it lacks the simple, simultaneous pairing of presence or absence between one antecedent and a corresponding result. Effectively, this method treats each 'situation' or experiment as one sample in a broader experiment demonstrating that whenever *a* is present, *Z* results, and whenever *a* is absent, *Z* is absent. The method makes the seemingly unreasonable assumption of 'all other things being equal'; yet this assumption is valid if the experiment is undertaken with adequate randomization.

Method of Concomitant Variations

If variation in an antecedent variable is associated systematically with variation in a consequent variable, then that antecedent causes the observed variations in the result. Symbolically, abc⇒Z, abΔc⇒ΔZ, ∴c⇒Z; or abc⇒WXYZ, abΔc⇒WXYΔZ, ∴c⇒Z.

The method of concomitant variations is like a combination of the methods of agreement and difference, but it is more powerful than either. Whereas the methods of agreement or difference merely establish an association, the method of concomitant variations quantitatively determines the relationship between causal and resultant variables. Thus the agreement and difference methods treat antecedents and consequents as attributes: either present or absent. The method of concomitant variations treats them as variables.

Usually one wants to know whether a relationship is present, and if so, what that relationship is. This method simultaneously addresses both questions. Furthermore, nonlinear relationships may fail the method of difference but be identified by the method of concomitant variation. For example, a method-of-difference test of the efficacy of a medication might find no difference between medicated and unmedicated subjects, because the medicine is only useful at higher dosages.

A quantitative relationship between antecedent and result, as revealed by the method of concomitant variation, may provide insight into the nature of that relationship. It also permits comparison of the relative importance of various causal parameters. This technique, however, is not immune to two limitations of the two previous methods:

- determination that a significant relationship exists does not prove causality; and
- other variables must be prevented from confounding the result. If they cannot be kept constant, then their potential biasing effect must be circumvented via randomization.

The correlation techniques described earlier in this chapter exploit the method of concomitant variations.

Method of Residues

If one or more antecedents are already known to cause part of a complex effect, then the other (residual) antecedents cause the residual part of the effect. Symbolically, abc⇒WXYZ, ab⇒WXY, ∴c⇒Z.

As defined restrictively above, this method is of little use because it assumes that every potentially relevant antecedent is being considered. Yet a pragmatic method of residues is the crux of much empirical science: identify the first-order causal relationship, then remove its dominating effect in order to investigate second-order and third-order patterns.

> The method of residues provided a decisive confirmation of Einstein's relativity: the theory accurately predicted Mercury's orbit, including the residual left unexplained by Newtonian mechanics. Another example is the discovery of Neptune, based on an analysis of the residual perturbations of the orbit of Uranus. Similarly, residual deviations in the orbits of Neptune and Uranus remain, suggesting the existence of a Planet X, which was sought unsuccessfully with Pioneer 10 and is still being looked for [Wilford, 1992b].

> The archaeological technique of sieving for potsherds and bone fragments is well known. Bonnichsen and Schneider [1995], however, have found that the fine residue is often rich in information: hair. Numerous animal species that visited the site or were consumed there can be identified. Human hair indicates approximate age of its donor and dietary ratio of meat to vegetable matter. Furthermore, it can be radiocarbon dated and may even have intact DNA.

* * *

The five inductive methods establish apparent causal links between variables or between attributes, but they are incomplete and virtually worthless without some indication of the confidence of the link. Confidence requires three ingredients:

- a quantitative or statistical measure of the strength of relationships, such as the correlation statistics described earlier in this chapter;
- discrimination between causal correlation and other sources of correlation, which is the subject of the next section; and
- an understanding of the power or confirmation value of the experiment, a subject that is discussed in Chapter 7.

The five inductive methods differ strikingly in confirmatory power. The Method of Difference and the Method of Concomitant Variations are the most potent, particularly when analyzed quantitatively with statistics. The Method of Agreement is generally unconvincing. Unfortunately, an individual hypothesis usually is not amenable to testing by all five methods, so one may have to settle for a less powerful test. Sometimes one can recast the hypothesis into a form compatible with a more compelling inductive test.

* * *

Correlation or Causality?

Causality needs correlation; correlation does not need causality. The challenge to scientists is to observe many correlations and to infer the few primary causalities.

Mannoia [1980] succinctly indicates how direct causal relationships are a small subset of all observed correlations. Observed statistical correlations (e.g., between *A* and *B*) may be:

• accidental correlations (1 of 20 random data comparisons is 'significant' at the 95% confidence level);

• two effects of a third variable that is causal and possibly unknown ($X \Rightarrow A$ & $X \Rightarrow B$);

• causally linked, but only indirectly through intervening factors ($A \Rightarrow X_1 \Rightarrow X_2 \Rightarrow B$, or $B \Rightarrow X_1 \Rightarrow X_2 \Rightarrow A$); or

• directly causally related ($A \Rightarrow B$ or $B \Rightarrow A$).

Earlier in this chapter, we examined quantitative measures of correlation strength and of the significance of correlations. Only an inductive conceptual model, however, can provide grounds for assigning an observed correlation to one of the four categories of causality/correlation. No quantitative proof is possible, and the quantitative statistical measures only provide clues.

Many factors affect or 'cause' change in a variable. Usually, our interest in these factors decreases with decreasing strength of correlation between the causal variables A_i and the effect B. In general, we judge the relative importance of various causal variables based on two factors: the strength of correlation and the rate of change dB/dA_i. High correlation strength means that much of the observed variation in effect B is somehow accounted for by variation in possible causal variable A_i. High rate of change means that a substantial change in effect B is associated with a modest change in causal variable A_i. However, rate of change alone can be misleading, for the total natural range of two causal variables A_1 and A_2 may be so different that dB/dA_1 could be larger than dB/dA_2 and yet A_2 causes more variation in B than A_1 does. Earlier in this chapter, we employed the correlation coefficient as a quantitative measure of correlation strength and the linear-regression slope as a measure of rate of change.

If one has three variables (C, D, and E) that are correlated, correlation strength can be used to infer likely relationships among them. Statistical techniques such as path analysis and analysis of covariance are best for determining these interconnections, but we will confine the present discussion to a more qualitative consideration of the problem. For example, suppose the correlation strengths among C, D, and E are as follows: C/D strong, D/E strong, and C/E weak. Probably, the weak relationship C/E is a byproduct of the two stronger correlations C/D and D/E, each of which may be causal. Direct causal connections ($A \Rightarrow B$) usually generate much stronger correlations than indirect ones ($A \Rightarrow X_1 \Rightarrow X_2 \Rightarrow B$). Extraneous factors affect each of the steps ($A \Rightarrow X_1$, $X_1 \Rightarrow X_2$, and $X_2 \Rightarrow B$) of the indirect correlation, thus weakening the overall correlation between A and B. Note, however, that relative strengths of correlations cannot establish causality; they only provide evidence about relative proximity of links among variables. For example, the pattern of C/D strong, D/E strong, and C/E weak could result either from $C \Rightarrow D \Rightarrow E$ or from $E \Rightarrow D \Rightarrow C$.

> Many surveys of U.S. voting patterns have shown that those who vote Republican have, on average, more education than Democratic voters. Does this mean that education instills Republican voting, or perhaps that higher intelligence inspires both greater education and Republican voting? Hoover [1988] uses this example to illustrate how social sciences need to beware of correlations induced by an unidentified third variable. More detailed and well-controlled surveys demonstrate that family wealth is the third variable: children of wealthier families tend to acquire a higher level of education and to be wealthier than average, and the voting pattern of wealthier individuals is more likely to be Republican than Democratic.

* * *

The following two examples illustrate the challenge of identifying the causality that manifests as correlation: the investigators had to design experiments to tease out this causal pattern. In both examples, epidemiological studies of a large population were used to identify a statistical association between a pair of variables.

> What is the effect of electromagnetic radiation on health? In one study, pregnant women who used video terminals more than 20 hours per week had twice as many miscarriages as did other kinds of female office workers. The authors of the study cautioned, however, that radiation was not necessarily the cause of this difference. For example, the video-intensive jobs might be more stressful.
>
> A statistical study of Denver children found that those who had lived near power-distribution lines were twice as likely to get cancer than other children. This study was criticized for its uncontrolled variables, so other investigators conducted a follow-up study designed to be much better controlled and more diagnostic. Contrary to the researchers' expectations, the new result was virtually the same as the original, so many scientists concluded that electromagnetic radiation really does seem to affect health. Note the origin of this change in opinions: the combination of a recognizably skeptical scientist and a tighter experiment [Stevens, 1992b].
>
> Compelling scientific evidence is required, because of the potentially staggering human and economic impacts if a causal link between electromagnetic radiation and health were confirmed. A synthesis of more than one hundred studies demonstrates that health impacts are generally negligible [Derry, 1999], but scientific concerns persist, particularly regarding possible long-term effects of cell phones.

> Is there a genetic predisposition to alcoholism? Research on this question exemplifies the problem of distinguishing between acquired and inherited characteristics. One of the most successful ways to attack such problems is by studying adopted children. For example, 30-40% of adopted children of alcoholics become alcoholics, compared to only 10% of the general population. This result constitutes good evidence for a genetic origin, but only because it was confined to children of alcoholic fathers; it is conceivable that an alcoholic mother could pass along an acquired dependence to her fetus, as occurs with heroin.
>
> In a different type of experiment, H. Begleiter found a much higher incidence of certain deficiencies in thinking and remembering among alcoholics than among non-alcoholics. Some of these deficiencies disappeared after the subjects stopped drinking, but others persisted for years. Was this evidence of permanent damage caused by alcohol? The author considered a radical alternative hypothesis: instead of the brain deficiency being caused by drinking, it preceded the drinking and was a trait among those most likely to become alcoholics. In studies of children, he found that 30-35% of the sons of alcoholic fathers had the deficiency, although only 1% of a control group did [Kolata, 1992a].

Rare scientists (e.g., Bauer, 1994) claim that the continuing debates about acquired vs. inherited characteristics illustrate deficiencies of sociology. Many non-scientists interpret the debates as revealing the fallibility of scientists. Instead, this research exemplifies the inductive ingenuity of those scientists who can recognize the possibility of a pattern among incredible complexity, then design a test that successfully isolates the primary variables.

Chapter 4: Deduction and Logic

"The supreme task of the physicist is to arrive at those universal elementary laws from which the cosmos can be built up by pure deduction. There is no logical path to these laws; only intuition, resting on sympathetic understanding, can lead to them." [Einstein, 1879-1955]

"'From a drop of water,' said [Sherlock Holmes], 'a logician could infer the possibility of an Atlantic or a Niagara without having seen or heard of one or the other. So all life is a great chain, the nature of which is known whenever we are shown a single link of it. Like all other arts, the Science of Deduction and Analysis is one which can only be acquired by long and patient study, nor is life long enough to allow any mortal to attain the highest possible perfection in it." [Doyle, 1893b]

[Harris, 1970]

* * *

Scientific deduction bears little similarity to the mythical conception conveyed by Sherlock Holmes. In science, obvious deductions are ubiquitous, insightful deductions are sporadic, and neither is infallible. We wield our logic with confidence, not noticing our occasional deductive errors. Before declaring that you are immune to such errors and skipping to the next chapter, please take ten minutes to attack the following problem:

> Imagine that four 3"x5" cards are on the table. You can see that each card has a single letter or number on its top: one has the letter 'A', one has 'B', one has the number '4', and one has the number '7'. You may assume that each card contains a single letter on one side and a single numeral on the other side. What cards is it necessary to turn over, to evaluate the validity of this rule: 'If a card has an A on one side, then it has a 4 on the other side'?

This problem, posed by Wason [1966], is considered by many to be a good example of the type of deductive decision-making that scientists face. Only 10% of college students answer the card problem correctly [Kuhn et al., 1988]. I suspect that you, like I, spent only a minute or two on the problem and got the wrong answer. Before proceeding, please consider the problem once more, this time actually using some props such as post-its, sheets of paper, or pencil and pad. Imagine that each card flip will be a major, time-consuming experiment. Will each experiment really be crucial to testing the hypothesis?

The correct answer to the card problem above is the two cards *A* and *7*. Many people answer *A* and *4*. The *B* card is clearly not useful, because it cannot prove or disprove the rule regardless of what is on the other side. Surprisingly, however, the same is true for the *4* card: even if it has an *A* on the other side, it supports but neither proves nor disproves the rule that any card with an *A* on one side has a *4* on the other side. In contrast, flipping the *7* card does test the rule, because the rule would be disproved if the other side is an *A*.

Many philosophers of science interpret the *A* & *4* answer as evidence of a confirmation bias: the chooser of the *4* card is seeking a result that confirms the hypothesis, rather than choosing the *7* card and potentially disproving the hypothesis. Scientists, in contrast, may justify choice of the *4* card as a search for patterns where they are most likely to be found. Not choosing the *7* card, however, is a failure to consider deductively the importance of potential results.

Two problems can involve identical deductive logic yet differ in difficulty. How a deductive problem is posed can affect the likelihood of correct results. Concrete examples are easier to solve than are the same problems expressed in symbols. For example, the success rate on the problem above was increased from 10% to 80% [Kuhn et al., 1988] when the problem was recast: given an envelope that may or may not be sealed and may or may not have a stamp on it, test the hypothesis, 'if an envelope is sealed, then it has a 5-pence stamp on it'.

Our greater facility with the concrete rather than with abstract deductions challenges the very basis of this decision-making. Possibly we do not even make decisions based on learned rules of formal logic [Cheng and Holyoak, 1985], but instead we recognize conceptual links to everyday experience [Kuhn et al., 1988]. The problem must seem real and plausible if there is to be a good chance of a successful solution; thus the postage problem is easier than the 4-card problem. In deductive logic, a similar strategy is often useful: recast the problem so that the logical structure is unchanged but the terms are transformed into more familiar ones. This technique, known as substitution, is one that we shall employ later in this chapter.

The four-card problem illustrates several points:

- prior thought can prevent needless experiments;
- sketches can be valuable in avoiding error;
- the same problem is more likely to be solved correctly if in familiar terms than if in abstract terms;
- confirmation bias is present in science, but to some extent it is a normal consequence of our pervasive search for patterns; and
- many people's 'deductive thinking' may actually be inductive pattern recognition of a familiar deductive form.

* * *

Logic

Logic means different things to different people. To Aristotle (384-322 B.C.), the 'Father of Logic', it was a suite of rules for deductive evaluation of syllogisms. To Peter Abelard (1079-1142) and William of Occam (1285-1349), Aristotelian logic was a useful launching point for development of a more comprehensive logic. G. W. Leibniz (1646-1716) sought to subsume all types of arguments within a system of symbolic logic. During the last century, symbolic logic has been the focus of so much study that it almost appeared to be the only type of logic. A notable exception was John Stuart Mill's Canons of inductive logic (Chapter 3).

Logic *is the science of argument evaluation*; it includes methods and criteria for deciding whether arguments are reliable. In this context, the term 'argument' has a meaning quite distinct from its everyday use as a difference of opinion: an **argument** *is a group of statements, consisting of evidence and a conclusion. Evidence statements are called premises,* and the conclusion is claimed to follow from these premises. For example, the following argument consists of three simplified statements, of which the first two are premises and the third is a conclusion:

All *A* are *B*.

All *B* are *C*.

Therefore, all *A* are *C*.

* * *

Deduction vs. Induction

Scientific logic has two distinctive branches: deduction and induction. Surprisingly, most scientists do not know the difference between these two types of inference. I, for example, used the word 'deduced' incorrectly in the title of my first major paper. Sherlock Holmes is indelibly associated with deduction, yet many of his 'deductions' were actually inductive interpretations based on subtle evidence.

To a first approximation, deduction is arguing from the general to the particular, whereas induction is arguing from the particular to the general [Medawer, 1969]. Often scientific induction does involve generalization from the behavior of a sample to that of a population, yet the following inductive argument goes from the general to the particular:

> In spite of many previous experiments, never has a relationship between variables *X* and *Y* been observed. Therefore, this experiment is unlikely to exhibit any relationship between *X* and *Y*.

In a deductive argument, the conclusion follows necessarily from the premises. In an inductive argument, the conclusion follows probably from the premises. Consequently, totally different standards are applied to deductive and inductive arguments. Deductive arguments are judged as valid or invalid by a black-or-white standard: *in a valid deductive argument, if the premises are true, then the conclusion must be true*. Inductive arguments are judged as strong or weak according to the likelihood that true premises imply a correct conclusion. Statistical arguments are always inductive. The following argument is inductively strong but deductively invalid:

> No one has ever lived more than 150 years.
>
> Therefore I will die before age 150.

A mnemonic aid for the difference between deduction and induction is: **de**duction is **de**finite; **in**duction is **in**definite and uncertain.

Both deductive and inductive arguments are evaluated in a two-step procedure:

- Does the conclusion follow from the premises?
- Are the premises true?

The order of attacking the two questions is arbitrary; usually one considers first whichever of the two appears to be dubious. The distinction between induction and deduction lies in the evaluation of whether the conclusion follows from the premises.

Here the focus is on deduction; induction was considered in Chapter 3. Before leaving the deduction/induction dichotomy, however, two common fallacies must be dispelled: 'scientific deduction is superior to induction,' and 'scientific induction is superior to deduction.' Three centuries ago, great minds battled over whether science *should be* deductive or inductive. René Descartes argued that science should be confined to the deductively certain, whereas Francis Bacon argued that the majority of scientific discoveries were empirical, inductive generalizations. A hallmark of the inception of rapid scientific progress was the realization that both deduction and induction are necessary aspects of science (Chapter 1). Yet the battle continues, fueled by misconceptions. For example, theoretical physicists such as Einstein probably would be outraged by the following statements from Beveridge's [1955] book on scientific methods:

> "Since deduction consists of applying general principles to further instances, it cannot lead us to new generalisations and so cannot give rise to major advances in science. On the other hand the inductive process is at the same time less trustworthy but more productive."

Inevitably, theoreticians value deduction and empiricists value induction, but the choice is based on taste rather than inherent superiority.

* * *

Scientific deduction uses the science of deduction, but the two do not share the same values or goals. Evaluating the validity of arguments is a primary objective of both, but scientific deduction places more emphasis on the premises. How can they be tested? Can the number of premises, or assumptions, be reduced, and if so what is the impact on the conclusion? How sensitive is the argument to the definition of terms in the premises? Are the premises themselves conclusions based on either deductive or inductive interpretation of other evidence?

Some scientists use a somewhat bootstrap logic that would be abhorrent to logicians. The technique is to tentatively assume an untested premise, and then see where it leads in conjunction with other, more established premises. If the resulting conclusion is one that is independently valued, perhaps on the basis of other deductive paths or perhaps on grounds of elegance or simplicity, then the premise may be tentatively accepted. These other standards of hypothesis evaluation are discussed more fully in Chapter 7.

* * *

Deductive Logic

Everyday language provides myriad opportunities for obscuring premises and conclusions, so the first step in evidence evaluation is usually the identification of premises and conclusion. Opinions, examples, descriptions, and many explanations are neither premise nor conclusion and are consequently not integral parts of an argument. Frequently, obvious premises are omitted from an argument:

> "Publish or perish" is an argument of the form:
>
> all A are B,
>
> not B,
>
> ∴ not A.
>
> Here we use the symbol '∴' to indicate 'therefore'. The premises are 'all successful scientists are paper publishers' and 'consider someone who is not a paper publisher'; the conclusion is 'that person is not a successful scientist'.

Premises may begin with one of the following flags: because, due to, since, given that, owing to, as indicated by, in that, . . . Likewise, most conclusions have an identifying flag: therefore, consequently, thus, accordingly, hence, so, as a result, it follows that, . . . Usually the conclusion is the first or last statement in an argument. Sometimes, however, one has to search for the conclusion by asking oneself, 'What is the author trying to convince me of?' For example, examine the following argument and identify the premises, conclusion, and any extraneous statements.

> Why should I have to study history? I am a scientist, I have more than enough to do already, I don't like history, and history is irrelevant to science.
>
> If one interprets the conclusion as 'History is irrelevant to me,' then the salient premises are 'History is irrelevant to scientists' and 'I am a scientist.' If one interprets the conclusion as 'History is a waste of time for me,' then the supporting premises are 'History is irrelevant to scientists,' 'I am a scientist,' and 'Doing history would prevent me from doing something more worthwhile.' The logic is valid, but some of the premises are dubious.

* * *

With deductive logic, each statement in the argument is either true or false. For the conclusion to be true, two critical preconditions must be met. First, the premises must be true. Second, the form of the argument must be valid. *A valid deductive argument is one in which the conclusion is necessarily true if the premises are true*. Validity or invalidity is totally independent of the correctness of the premises; it depends only on the form of the argument -- thus the term **form**al logic.

The following arguments demonstrate the distinction between the roles of premises and of logical form in determining the correctness of a conclusion:

All dogs are cats. All cats are animals. Therefore, all dogs are animals.	Valid form, but one false premise, so the argument is incorrect (although the conclusion happens to be true).
All dogs are mammals. All mammals are animals. Therefore, all dogs are animals.	Valid form, true premises, so the argument is correct and the conclusion must be true.
All dogs are mammals. All cats are mammals. Therefore, all dogs are cats.	True premises, but invalid form, so the argument is invalid and does not yield this conclusion.

For these three examples, the reader already knows which conclusions are true and which are false without even evaluating the arguments. For scientific arguments, however, it is crucial that one considers separately the two elements -- premise correctness and argument form -- rather than accept or reject the argument based on whether or not the conclusion sounds right. Evaluation of premises requires subjective judgment based on local expertise. Evaluation of argument form, in contrast, is objective. With some practice and a few guidelines, the reader can avoid using invalid argument forms and recognize them in publications. Such is the main goal of this chapter.

* * *

Classification Statements

A building block of deductive logic is the classification statement; logicians use the term categorical proposition. The classification statement consists of a subject and predicate, and it states that

members of the subject category are or are not included in the predicate category. For example, the statement 'all scientists are people' is a classification statement, in which '**s**cientists' is the **s**ubject and '**p**eople' is the **p**redicate. The four types of classification statement are:

- *All S are P:* The entire subject class lies within the predicate class. Every member of the subject class is also a member of the predicate class.
- *No S are P:* The entire subject class is excluded from, or outside, the predicate class. No member of the subject class is a member of the predicate class.
- *Some S are P:* At least one member of the subject class lies within, and is a member of, the predicate class.
- *Some S are not P:* At least one member of the subject class lies outside, and is not a member of, the predicate class.

Note that 'some' means at least one; it does not mean 'less than all'. Thus it is possible for both statements 'All *S* are *P*' and 'Some *S* are *P*' to be true for the same *S* and *P*; if so, the former statement is more powerful. Similarly, both statements 'Some *S* are *P*' and 'Some *S* are not *P*' may be true for the same *S* and *P*.

The statements 'All *S* are *P*' and 'No *S* are *P*' are sometimes referred to as universal statements because they apply to every member of a class. In contrast, the statements 'Some *S* are *P*' and 'Some *S* are not *P*' apply not to every member but instead to a particular subset; thus they are referred to as particular statements.

Deductive Aids: Venn Diagrams and Substitution

The four classification statements can be illustrated diagrammatically as shown in Figure 17.

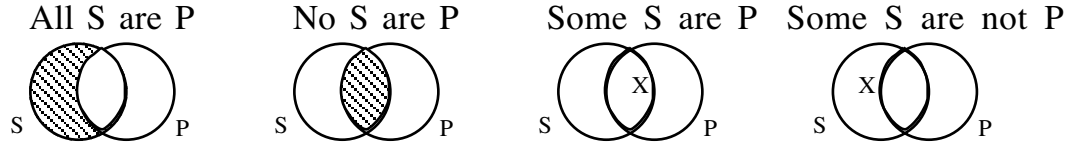


Figure 17. Classification statements, expressed as Venn diagrams.

John Venn, a 19th-century logician, invented this technique of representing the relationship between classes. Each class is represented by a circle; in this case there are only the two classes *S* or *P*. Potential members of the class are within the circle and individuals not belonging to the class are outside the circle. The overlap zone, lying within both circles, represents potential members of both classes. Hatching indicates that a zone contains no members (mathematics texts often use exactly the opposite convention). An *X* indicates that a zone contains at least one ('some') member. Zones that contain neither hatching nor an *X* may or may not contain members. In the next section, we will observe the substantial power of Venn diagrams for enhancing visualization of deductive statements or arguments. For now, it suffices to understand the Venn representations above of the four classification statements:

- *All S are P*: The zone of *S* that is not also *P* is empty (hatched), and the only possible locations of *S* are in the zone that overlaps *P*. Ergo, all *S* are *P*.
- *No S are P*: The zone of *S* that overlaps *P*, *i.e.* that is also *P*, is empty.
- *Some S are P*: The *X* indicates that at least one member lies within the zone that represents members of both *S* and *P*. The remaining members of *S* or *P* may or may not lie within this zone.

- *Some S are not P*: The *X* indicates that at least one member lies within the zone that represents members of *S* but not of *P*. Other members of *S* may or may not lie within *P*.

* * *

Substitution is a powerful technique for recognizing valid and invalid deductive arguments. Validity depends only on the form of the argument. Therefore, we can replace any arcane or confusing terms in a deductive argument with familiar terms, then decide whether or not the argument is valid. For example, the following four arguments all have the same invalid form:

> If a star is not a quasar, then it is theoretically impossible for it to be any type of star other than a neutron star. This follows from the fact that no neutron stars are quasars.
>
> No neutron stars are quasars. Therefore, no non-quasars are non-neutron stars.
>
> No S are P. ∴ no non-P are non-S
>
> No cats are dogs. Therefore, no non-dogs are non-cats.

Recognizing that the first three arguments are invalid is easy for some readers and difficult for others. Some of us experience mind-glaze when faced with arguments involving unfamiliar and highly technical terms; others find abstract, symbolic notation even more obscure. Some can analyze arguments easier when the argument is in a standard notation; others prefer their arguments to be couched in everyday language. Everyone can immediately recognize the fallacy of the cats-and-dogs argument, for obviously the world is full of objects that are neither cat nor dog. If this cats-and-dogs argument is invalid, then the other three arguments *must* be invalid because they have the same form.

Substitution relies on four principles that we have encountered in this chapter:

- Validity or invalidity of a deductive argument depends only on the form of the argument, not on its topic (*note*: this is not true for inductive arguments).
- A valid deductive argument is one in which the conclusion is necessarily true if the premises are true (*note*: this is not true for inductive arguments).
- If we know that the premises of an argument are true and yet the conclusion is false, then the argument *must* be invalid.
- Validity or invalidity is much easier to recognize for arguments about familiar objects than for abstract arguments.

To employ substitution, simply identify the elements of the argument and replace each element with a familiar term. In the examples above, the elements are neutron stars and quasars, or *S* and *P*, or cats and dogs, and the structural equivalents are *S*=neutron stars=cats and *P*=quasars=dogs. Formal logic assumes that the premises are true, so it is easiest if one picks substitutions that yield a true initial statement. Then, an absurd result can be attributed correctly to invalid logic.

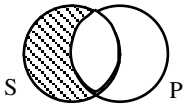
Substitution may be the main way that most people (logicians excluded) evaluate deductions, but this method seldom is employed consciously. Instead, we unconsciously perceive that an argument is familiar, because it is similar in form to arguments that we use almost every day. Conversely, we may recognize that an argument sounds dubious, because it seems like a distortion of a familiar argument form. With that recognition, we then can deliberately employ substitution to test the argument.

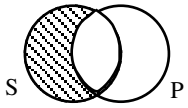
* * *

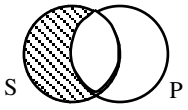
Logically Equivalent Statements

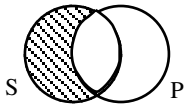
Venn diagrams permit us to identify or remember **logically equivalent statements**. Such statements have exactly the same truth value (whether true or false) as the original. The Venn diagrams in Figure 18 permit us to identify which apparent equivalences are valid (identical Venn diagrams) and which are invalid (different Venn diagrams).

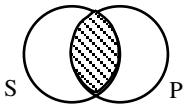
Valid equivalent statements:

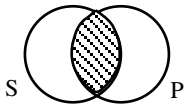
All S are P: 

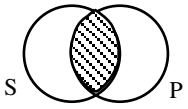
No S are non-P: 

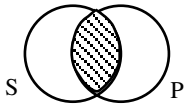
All S are P: 

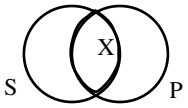
All non-P are non-S: 

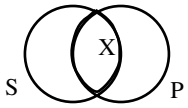
No S are P: 

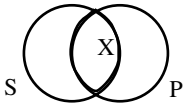
No P are S: 

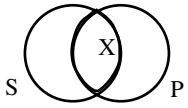
No S are P: 

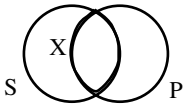
All S are non-P: 

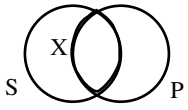
Some S are P: 

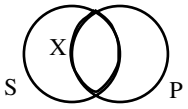
Some P are S: 

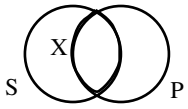
Some S are P: 

Some S are not non-P: 

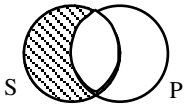
Some S are not P: 

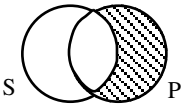
Some S are non-P: 

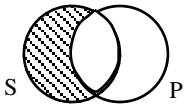
Some S are not P: 

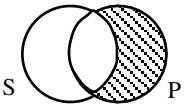
Some non-P are not non-S: 

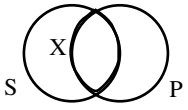
Superficially similar but non-equivalent statements:

All S are P: 

No P are non-S: 

All S are P: 

All P are S: 

Some S are not P: 

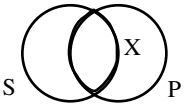
Some P are not S: 

Figure 18. Valid and invalid equivalent statements, and their Venn diagrams.

Logicians use the terms conversion, obversion, and contraposition to define three types of logically equivalent statements, but we will not need to memorize these terms. Below are listed on the right the only logically equivalent statements to those on the left:

Initial statement	Logically equivalent statements	
All S are P.	No S are non-P.	All non-P are non-S.
No S are P.	No P are S.	All S are non-P.
Some S are P.	Some P are S.	Some S are not non-P.
Some S are not P.	Some S are non-P.	Some non-P are not non-S.

Some logically equivalent statements seem cumbersome and overloaded with negatives. That apparent weakness is a strength of the concept of logical equivalence, for we may encounter a statement on the right and want to translate it into a familiar classification statement.

The concept of logical equivalence can also be useful in experimental design. For example, it might be impossible to show that 'some *S* are *P*' but easy to show that 'some *P* are *S*'. In Chapter 7 we will consider the Raven's Paradox: the two statements 'All ravens are black' and 'All non-black things are non-ravens' may be logically equivalent, but testing the latter would involve an inventory of the universe.

* * *

For recognizing logically equivalent statements, substitution is an alternative to Venn diagrams. For example, replace *S* with **s**cientists and replace *P* with either **p**eople, **p**hysicists, or **p**oliticians, whichever gives a true initial statement:

Valid equivalent statements:

All Scientists are People.	No Scientists are non-People.
All Scientists are People.	All non-People are non-Scientists.
No Scientists are Politicians.	No Politicians are Scientists.
No Scientists are Politicians.	All Scientists are non-Politicians.
Some Scientists are Physicists.	Some Physicists are Scientists.
Some Scientists are Physicists.	Some Scientists are not non-Physicists.
Some Scientists are not Physicists.	Some Scientists are non-Physicists.
Some Scientists are not Physicists.	Some non-Physicists are not non-Scientists.

Non-equivalent statements:

All Scientists are People.	No People are non-Scientists.
All Scientists are People.	All People are Scientists.
Some Scientists are not Physicists.	Some Physicists are not Scientists.

* * *

Relationships Among Statements

The four types of classification statement are formally related in truth value, regardless of the subjects of the statements. The relationships can be summarized in what is called the **square of opposition** (Figure 19).

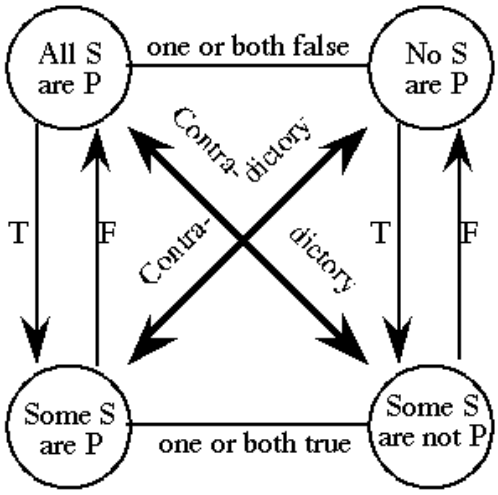


Figure 19. Square of opposition.

The strongest relationship among the statements is that of contradiction along the diagonals: if a statement is true, then its diagonal is false, and vice versa. Without even substituting familiar terms for the subject and predicate, one can recognize readily that:

• 'All *S* are P' contradicts the statement 'Some *S* are not P', and

• 'No *S* are P' contradicts the statement 'Some *S* are P'.

Horizontally along the top, one or both of the statements invariably is false:

- If 'All *S* are *P*' is true, then 'No *S* are *P*' must be false;
- If 'No *S* are *P*' is true, then 'All *S* are *P*' must be false;
- If either 'All *S* are *P*' or 'No *S* are *P*' is false, we cannot infer that the other statement is true; possibly both are false and 'Some *S* are *P*'.

Horizontally along the bottom, one or both of the statements invariably is true:

- If 'Some *S* are *P*' is false, then 'Some *S* are not *P*' must be true;
- If 'Some *S* are not *P*' is false, then 'Some *S* are *P*' must be true;
- Both statements may be true: some *S* are *P* while other *S* are not *P*.

Vertically, the statements lack the perfect symmetry that we saw diagonally and horizontally. Instead, imagine truth flowing downward (from the general to the particular) and falsity flowing upward (from the particular to the general):

• If 'All *S* are *P*' is true, then it is also true that 'Some *S* are *P*'.

The knowledge that 'All *S* are *P*' is false, however, does not constrain whether or not 'Some *S* are *P*'.

• Similarly, if 'No *S* are *P*' is true, then it is also true that 'Some *S* are not *P*'. The knowledge that 'No *S* are *P*' is false, however, does not constrain whether or not 'Some *S* are not *P*'.

• If 'Some *S* are *P*' is false, then 'All *S* are *P*' must also be false. The knowledge that 'Some *S* are *P*' is true, however, does not indicate whether or not 'All *S* are *P*'.

• Similarly, if 'Some *S* are not *P*' is false, then 'No *S* are *P*' must also be false. The knowledge that 'Some *S* are not *P*' is true, however, does not indicate whether or not 'No *S* are *P*'.

These relationships can be visualized more easily with a square of opposition composed of Venn representations of the four types of statement (Figure 20).

For example, the Venn diagrams demonstrate the incompatible, contradictory nature of diagonal statements such as 'All *S* are *P*' and 'Some *S* are not *P*'.

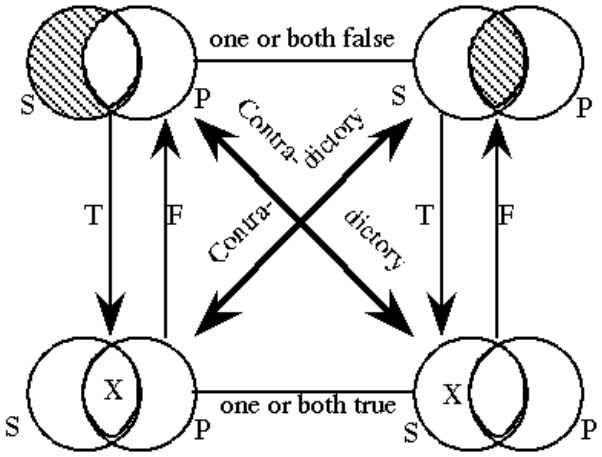


Figure 20. Venn square of opposition.

Table 8 summarizes the relationships that can be determined between any two of the classification statements by examination of the square of opposition.

Table 8. Relationships among classification statements.

		All S are P	No S are P	Some S are P	Some S are not P
If 'All S are P' true,	then		false	true	false
If 'All S are P' false,	then		unknown	unknown	true
If 'No S are P' true,	then	false		false	true
If 'No S are P' false,	then	unknown		true	unknown
If 'Some S are P' true,	then	unknown	false		unknown
If 'Some S are P' false,	then	false	true		true
If 'Some S are not P' true,	then	false	unknown	unknown	
If 'Some S are not P' false,	then	true	false	true	

Finally and most simply (for me at least), one can immediately see the impact of any one statement's truth value on the other three statements through substitution. Again I substitute Scientist for S, and either People, Physicists, or Politicians for P, whichever fits the first statement correctly. For example, if I assume (correctly) that 'Some scientists are physicists' is true, then 'No scientists are physicists' must be false, and I need additional information to say whether 'All scientists are physicists' or 'Some scientists are not physicists'. Some caution is needed to assure that my conclusions are based on the evidence rather than on my independent knowledge. For example, I know that 'All scientists are physicists' is false but I cannot infer so from the statement above that 'Some scientists are physicists'. As another example, if I assume (naïvely) that 'Some scientists are politicians' is false, then it also must be true that 'No scientists are politicians' and that 'Some scientists are not politicians'. Furthermore, the statement that 'All scientists are politicians' must be false.

* * *

Syllogisms

Syllogism is the deductive solution of a pervasive scientific problem: what is the relationship between the two classes *A* and *C*, given that I know the relation of both *A* and *C* to the third class *B*?

Aristotle loved syllogisms. He systematized them, developed rules for and patterns among them, and promoted them as the foremost tool for analysis of arguments. But what is a syllogism? Let us examine the syllogism using Aristotle's own example:

> All men are mortal.
>
> Socrates is a man.
>
> Therefore Socrates is mortal.

This argument is recognizable as a syllogism by these characteristics:

- the argument consists of three statements;
- two of the statements (in this case the first and second) are premises and the third is a conclusion that is claimed to follow from the premises.

In so-called standard form such as the Socrates syllogism, the third statement is the conclusion, containing a subject ('Socrates') and predicate ('mortal'), the first statement is a premise dealing with the predicate, and the second statement is a premise dealing with the subject.

Syllogisms are of three types: categorical, hypothetical, and disjunctive. We will consider hypothetical syllogisms briefly later in this chapter. The Socrates syllogism is categorical: three classification statements, each beginning explicitly or implicitly with one of the three words 'all', 'no', or 'some', with two terms in each statement, and with each term used a total of twice in the argument. Each term must be used in exactly the same sense both times. For example, *man* cannot refer to *mankind* in one use and *males* in the second; this is the fallacy of equivocation, described in a later section.

Chambliss [1954] succinctly comments:

> "The syllogism does not discover truth; it merely clarifies, extends, and gives precision to ideas accepted as true. It is, according to Aristotle, 'a mental process in which certain facts being assumed something else differing from these facts results in virtue of them.'"

Aristotle's description that "something else differing from these facts results" is a bit misleading in its hint of getting something for nothing. The conclusion does not really transcend the premises; instead it is really immanent, an implication of the premises that may or may not be obvious. Rather than discover truth, the syllogism reveals the implications of our assumptions. As such, it is a fundamental step in the hypothetico-deductive method (better known as *the* scientific method).

Syllogisms can be difficult to recognize in everyday language. Formal analysis of syllogistic logic requires a translation from everyday language into the so-called standard syllogism form. This translation may involve reorganizing the statements, recognizing that a term can be much longer than one word, using logical equivalences to reduce terms, supplying an omitted (but implied) premise or conclusion, or breaking apart a compound argument into its component syllogisms. This translation is useful to learn but beyond the scope of this book; the reader is encouraged to consult a textbook on logic and practice translation of the many examples therein. Here we focus on the analysis of standard-form syllogisms, because familiarity with standard-form syllogisms has a fringe benefit: invalid syllogisms will sound dubious and invite closer scrutiny, even if they are couched in everyday language.

* * *

Categorical Syllogisms

Categorical syllogisms have 256 varieties; only 24 are valid. Any one of these 256 can occur in scientific arguments or everyday life, and we should be able to recognize whether it is valid or invalid. Simply but brutally put, we cannot always avoid false assumptions, false inductions, or misleading data, but *we must avoid invalid deductions*. A scientist who incorrectly judges the validity of a syllogism may design and undertake an entire experiment based on a fallacious expectation of its potential meaning.

Table 9: Valid categorical syllogisms [Hurley, 1985].

Unconditionally valid:

All M are P.	All S are M.	∴ All S are P.	
No M are P.	All S are M.	∴ No S are P.	
All M are P.	Some S are M.	∴ Some S are P.	
No M are P.	Some S are M.	∴ Some S are not P.	
No P are M.	All S are M.	∴ No S are P.	
All P are M.	No S are M.	∴ No S are P.	
No P are M.	Some S are M.	∴ Some S are not P.	
All P are M.	Some S are not M.	∴ Some S are not P.	
Some M are P.	All M are S.	∴ Some S are P.	
All M are P.	Some M are S.	∴ Some S are P.	
Some M are not P.	All M are S.	∴ Some S are not P.	
No M are P.	Some M are S.	∴ Some S are not P.	
All P are M.	No M are S.	∴ No S are P.	
Some P are M.	All M are S.	∴ Some S are P.	
No P are M.	Some M are S.	∴ Some S are not P.	

Conditionally valid:

All M are P.	All S are M.	∴ Some S are P.	(S must exist)
No M are P.	All S are M.	∴ Some S are not P.	(S must exist)
All P are M.	No S are M.	∴ Some S are not P.	(S must exist)
No P are M.	All S are M.	∴ Some S are not P.	(S must exist)
All P are M.	No M are S.	∴ Some S are not P.	(S must exist)
All M are P.	All M are S.	∴ Some S are P.	(M must exist)
No M are P.	All M are S.	∴ Some S are not P.	(M must exist)
No P are M.	All M are S.	∴ Some S are not P.	(M must exist)
All P are M.	All M are S.	∴ Some S are P.	(P must exist)

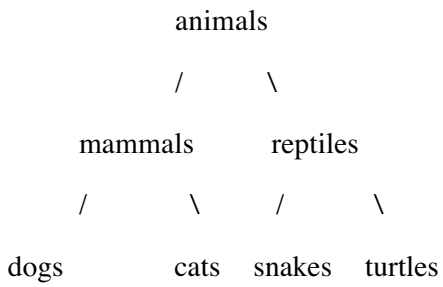
Many strategies *could* be employed to distinguish between valid and invalid categorical syllogisms:

- random choice (not a very scientific basis for decision-making at any time, but particularly when the chance of winning is only 24/256);
- memorization, an old, laborious standby;
- knowing where the answer can be found (Table 9);
- recognition that the correct solutions all obey a few rules (only five rules are needed for successful separation of the 24 valid syllogisms from the 232 invalid ones);
- sketching Venn diagrams;
- substitution, in which we recognize that the problem structure is identical to one whose answer is known.

All except for the 'random choice' option are acceptable solutions to the problem, but memorization and substitution have the strong advantage of much greater speed. In the remainder of this section, I list the valid syllogisms for easy reference, and then I describe substitution -- the easiest closed-book technique for evaluating syllogisms.

* * *

Substitution is an easy way to evaluate categorical syllogisms. As with the evaluation of any formal logic, the validity of the form is independent of the actual terms used. If we insert familiar terms into the syllogism, choosing ones that yield true premises, then an untrue conclusion *must* indicate an invalid syllogism. For evaluation of categorical syllogisms, I select substitutions from the following classification tree:



The danger of substitution is that a true conclusion does not prove that the logic is valid, as we saw above for the syllogism "Some mammals are dogs; some mammals are cats; therefore no cats are dogs." Substitution can prove that an argument is invalid but, unfortunately, cannot prove that it is valid. If the premises are true, a substitution that yields a true conclusion may or may not be of valid form. In contrast, a substitution with true premises and false conclusion must be of invalid form. Thus one needs to consider several substitutions, to see whether any case can prove invalidity. For example, the following argument is not disproved by the first substitution but is disproved by the second one:

Some physicists are theoreticians.
Some astronomers are theoreticians.
Therefore some physicists are astronomers.

Some dogs are animals.
Some mammals are animals.
Therefore some dogs are mammals.

Some dogs are mammals.
Some cats are mammals.
Therefore some dogs are cats.

Usually, an invalid syllogism couched in familiar terms feels wrong, even if the conclusion is true. Further brief thought then generates a variant that proves its invalidity. Using the 'animal tree' to test syllogisms can generally avoid the juxtaposition of invalid logic and true conclusion: simply confine each statement to adjacent levels in the animal tree, rather than creating statements like 'some dogs are animals' that skip a level.

* * *

Hypothetical Syllogisms

Like categorical syllogisms, hypothetical syllogisms consist of two premises and a conclusion. Unlike categorical syllogisms, one or both of the premises in a hypothetical syllogism is a conditional statement: 'if A, then B'.

We can express a conditional, or if/then, statement symbolically as A⇒B. The statement A⇒B can be read as 'A implies B' or as 'if A, then B'; the two are logically equivalent. Both statements state that A is a necessary and sufficient condition for B.

If both premises in a hypothetical syllogism are if/then statements, then only three forms of syllogism are possible:

Valid	Invalid	Invalid
S⇒M.	S⇒M.	M⇒S.
M⇒P.	P⇒M.	M⇒P.
∴ S⇒P.	∴ S⇒P.	∴ S⇒P.

Another type of hypothetical syllogism has one if/then statement, a statement that one of the two conditions is present or absent, and a conclusion about whether the other condition is present or absent. Symbolically, we can indicate presence (or truth) by *S* or *P*, and absence by *-S* or *-P*. If only one premise is an if/then statement, two valid and two invalid forms of syllogism are possible:

Valid	Invalid	Invalid	Valid
S⇒P	S⇒P	S⇒P	S⇒P
S	-S	P	-P
∴ P	∴ -P	∴ S	∴ -S

As with categorical syllogisms, hypothetical syllogisms are readily testable through substitution. The substitution that I use treats *if/then* as a mnemonic for 'if the hen':

A: if the hen lays an egg;
B: we cook omelettes;
C: we eat omelettes.

This substitution readily distinguishes invalid from valid hypothetical syllogisms:

Valid: A⇒B. If the hen lays an egg, then we cook omelettes.
B⇒C. If we cook omelettes, then we eat omelettes.
∴ A⇒C. Therefore, if the hen lays an egg, we eat omelettes.

Invalid: A⇒B. If the hen lays an egg, then we cook omelettes.
C⇒B. If we eat omelettes, then we cook omelettes.
∴ A⇒C. Therefore, if the hen lays an egg, we eat omelettes (invalid; eating omelettes is not necessarily related to the hen's laying).

Invalid: B⇒A. If we cook omelettes, then the hen lays an egg.
B⇒C. If we cook omelettes, then we eat omelettes.
∴ A⇒C. Therefore, if the hen lays an egg, we eat omelettes (invalid, not because the first premise is absurd but because the hen's laying and our omelette eating are not necessarily related).

Valid: A⇒B. If the hen lays an egg, then we cook omelettes.
A. The hen laid an egg.
∴ B. Therefore, we cook omelettes.

Valid: A⇒B. If the hen lays an egg, then we cook omelettes.
-B. We are not cooking omelettes.
∴ -A. Therefore, the hen did not lay an egg.

Invalid: A⇒B. If the hen lays an egg, then we cook omelettes.
-A. The hen did not lay an egg.
∴ B. Therefore, we are not cooking omelettes. (invalid; maybe we can get eggs elsewhere)

Invalid: A⇒B. If the hen lays an egg, then we cook omelettes.
B. We are cooking omelettes.
∴ A. Therefore, the hen laid an egg. (invalid; maybe we can get eggs elsewhere)

The last two fallacies above are so obviously wrong that we might dismiss them as irrelevant to scientists. When couched in technical terms, however, these invalid syllogisms do appear occasionally in print. Both fallacies imply confusion between necessary and sufficient conditions. Both are deductively invalid, but they may have some inductive validity:

Valid: If the hen lays an egg, then we cook omelettes.
The hen did not lay an egg.
Therefore, we *may* not cook omelettes.
(the hen's failure is a setback to our omelette plans, but maybe we can get eggs elsewhere)

Valid: If the hen lays an egg, then we cook omelettes.
We are cooking omelettes.
Therefore, the hen *may* have laid an egg. (true, but maybe we got eggs elsewhere)

This second hypothetical syllogism is a cornerstone of **scientific induction**: "If hypothesis (*H*) entails Evidence (*E*), and *E* is true, then *H* is probably true." It is fallacious to conclude that H is definitely true, but the evidence is relevant to evaluation of the hypothesis.

* * *

Pitfalls: Fallacious Arguments

After a bit of practice, one can readily recognize syllogistic arguments that are expressed in ordinary language, and one can evaluate them by examining their structures. Many arguments can appear to be structurally valid and yet be fallacious; such arguments yield a false conclusion even if the premises are true. These fallacies exhibit an error in execution, such as subtle problems in their premises, use of apparently relevant but logically irrelevant evidence, an incorrect connection of premises to conclusion, and grammatical errors or ambiguities. Many of these fallacies are genuine

pitfalls to scientists. Most are deductive pitfalls, but a couple of inductive pitfalls (e.g., hasty generalization) are included here because of their similarity to deductive pitfalls.

The list of fallacies that follows is loosely based on the compilation of Hurley [1985]. Other logicians lump or split these fallacies differently and describe them with different jargon. For our purposes, the names applied to these fallacies have limited usefulness; instead, our goal is to recognize when an argument is fallacious. Practice with a variety of examples is the key, and logic textbooks have a wealth of examples.

Most fallacies fall into one of four types: problems in a premise, extraneous extra evidence, faulty link between premises and conclusion, or case-dependent relationship between parts and whole. Table 10 gives an overview of these different kinds of fallacy, and the remainder of this chapter examines these fallacies in more detail.

* * *

Table 10. Varieties of fallacious argument.

Problems in a premise:

Fallacy	**Premises**	**other 'evidence'**	**⇒**	**Conclusion**
false dichotomy	2 choices assumed	other choices omitted		
suppressed evidence	weakness ignored			
ambiguity	ambiguity		misinterpreted	
false cause	noncausal, yet assumed causal			
slippery slope	unlikely chain of events		flawed links	

Extraneous other evidence:

Fallacy	**Premises**	**other 'evidence'**	**⇒**	**Conclusion**
appeal to authority		experts say . . .		
personal attack		fools say . . .		
mob appeal		rest of group says . . .		
might makes right		accept or suffer consequences		
extenuating circumstances		extenuating circumstances		
red herring		smoke-screen distraction		

Faulty link between premises and conclusion:

Fallacy	Premises	other 'evidence'	⇒	Conclusion
missing the point	imply conclusion A			conclusion B drawn
overinterpreting	uncertain			definite
begging the question #1		dubious premise ignored		
begging the question #2	validated by conclusion		circular reasoning	validated by premises
equivocation	one meaning for key word			another meaning for same word
straw man			tested with bad example	

Case-dependent relationship between parts and whole:

Fallacy	Premises	other 'evidence'	⇒	Conclusion
false extrapolation to whole	parts		attribute misapplied	whole
false extrapolation to parts	whole		attribute misapplied	part
false extrapolation to individual	general		attribute misapplied	individual
hasty generalization	nonrepresentative individual		generalized	general

* * *

Fallacies Resulting from Problems in a Premise

> For scientists, few 'victimless' crimes are as outrageous as the burning of the Alexandria library, and with it the destruction of so much ancient knowledge and culture. One legend is that when the Muslim Amrou Ibn el-Ass captured Alexandria, he sought his caliph's guidance on the fate of the library. Caliph Omar responded that the library's books are either inconsistent or consistent with the Koran. If inconsistent, they are heretical; if consistent, they are redundant. In either case they should be burned. [Gould, 1990]
>
> The story is apocryphal and, I suspect, wrong. The library was probably destroyed in 389 A.D., not 642 A.D., and the Muslims embraced other cultures and their science at a time when Christians were suppressing them. As a memorable example of false dichotomy, however, the story is unsurpassed.

A valid deduction does not imply a correct conclusion; accurate premises or assumptions are also essential. When reading a research paper, the scientist must seek and evaluate the premises. Incorrect or overlooked premises are probably the dominant source of incorrect scientific deductions, and these errors can take several forms:

- **False dichotomy** is an incorrectly exclusive 'either . . .or. . .' statement in one of the premises. When one choice is eliminated by another premise, the other choice is accepted incorrectly as the conclusion. The logic is valid, and if there truly are only two choices then the conclusion is valid:

> "Either you subscribe to the journal or you don't. Your subscription lapsed, and therefore you don't subscribe to the journal."

The fallacy of false dichotomy is that the either/or premise is false if more than two choices exist. Therefore the conclusion is invalid:

> "Either the hypothesis is proved or disproved. This experiment did not prove the hypothesis. Therefore it must have disproved it." Unfortunately, science is almost always less efficient than this. Experiments may support hypotheses, refute them, or disprove them, but never prove them.

False dichotomy is frequent among the general public.

Sometimes one premise and the conclusion are obvious and unstated:

> "Either make at least 100 measurements or skip the experiment entirely." The premises (*P*) and conclusion (*C*) are: *P1*: the experiment is worthless if <100 measurements are made; *P2*: surely you want the experiment to be worthwhile; and *C*: therefore you will want to do at least 100 measurements.

• **Suppressed evidence** is the omission of evidence that weakens or fatally undermines one premise. This fallacy is frequent among both lay people and scientists. Few scientists deliberately hide an assumption. Instead, they may suppress evidence passively, by an unconscious 'forgetting' or by a conscious decision that the evidence is too flawed to warrant mention. A different, but related, lapse of objectivity is the ignoring of evidence that leads to a competing conclusion.

• **Ambiguity** creates a fallacious argument, when misinterpretation of an ambiguous premise results in a wrong conclusion. Usually the ambiguity arises from punctuation or grammar and is merely a temporary distraction while reading a publication:

> "We analyzed our experiments on monkeys using multivariate statistics." Smart monkeys!

Misinterpretation of someone else's ambiguously stated premise is more serious. People often are unaware of ambiguities in their own statements, because of familiarity with the subject. Others then misinterpret the statement, leading them to incorporate it into an argument that is doomed by the incorrect premise.

> A sign on a beach says, "Sharks! No swimming!" [Ennis, 1969]

> My colleagues and I have often succumbed to the fallacy of ambiguity in interpreting telexes. The sender cannot foresee the ambiguity that cost-saving brevity has introduced. For example: ". . . STOP MISS YOU STOP LOVE END"

• **False cause** is an argument in which a relationship is incorrectly assumed to be causal. Several types of associations can be misinterpreted as causal: (1) one event may precede another and become misidentified as its cause; (2) the cause may be confused with the effect if the two are nearly simultaneous; (3) a variable may control two others and thereby give those two an indirect association; and (4) the apparent association may be coincidental. Determining causality and dodging the potential pitfall of false cause are fundamental aspects of science. They are discussed in more detail in Chapter 3.

• **Slippery slope** is an argument in which the premises form a chain reaction of assumed causal consequences, beginning with some initial event and culminating with a conclusion. One step onto a slippery slope causes one to slide all the way to an undesirable outcome. The arguer's purpose is usually to prevent that first step. The slippery-slope fallacy is the invalid assumption that a full chain reaction invariably follows the initial event. Almost all chain reactions are invalid, because each step requires a causality that is both necessary and sufficient; only then are alternative paths precluded. Thus chain-reaction arguments are particularly vulnerable to the fallacy of false cause.

Slippery-slope logic is used with mixed success by many fundamentalist preachers. Seldom is it used in science, but sometimes the link between a hypothesis and a testable prediction can involve several steps. If so, one must evaluate whether each step validly involves either pure deduction or a necessary and sufficient causality.

> The most familiar example of a slippery slope, at least to those in my age group, is domino theory. Used successfully in the early justifications of the Vietnam war, domino theory said that if Vietnam were to fall to communism, through chain reaction all of Southeast Asia would eventually become communist. Domino theory was wrong.

> In attempting to refute Galileo's claim that he had discovered satellites of Jupiter, astronomer Francesco Sizi [Holton and Roller, 1958] used a *slippery-slope* argument:
>
> "The satellites are invisible to the naked eye and therefore can have no influence on the earth and therefore would be useless and therefore do not exist."

* * *

Fallacies Employing Extraneous Other Evidence

> When ego is involved, scientific arguments can get personal. This was often the case for Isaac Newton, as the following letter [~1700] illustrates. Note that Newton attempts to demolish an idea without giving a single shred of evidence:
>
> "That gravity should be innate, inherent and essential to matter, so that one body may act upon another at a distance through a *vacuum*, without the mediation of any thing else, by and through which their action and force may be conveyed from one to another, is to me so great an absurdity, that I believe no man who has in philosophical matters a competent faculty of thinking, can ever fall into it."

Unlike Newton's argument, most arguments do involve evidence that can be evaluated in terms of premises and deductive or inductive conclusions. They may also, however, contain a collage of other information that the proponent considers to be relevant but that is extraneous to the core deductive argument. Often this extraneous information is emotionally charged, and the evaluator must cull the deductive argument from among the distractions.

• **Appeal to authority** is the claim that an argument should be accepted because some expert accepts it. Ideally, scientists do not appeal to authority; they evaluate evidence personally. In practice, however, we limit such analyses primarily to our own field, and we tentatively accept the prevailing wisdom of scientists in other fields. The appeal to authority must be considered pragmatically, based on how much more experience the 'authority' has than the arguers have, how mainstream the authority's view is, and how feasible it is for the arguers to evaluate all of the evidence.

> For example, when a biologist considers a physics argument, it is valid to give weight to what physicists believe. Yet when a physicist considers a physics argument, it is a fallacy to accept it merely because some 'great' physicist believes it.

• **Personal attack** is a criticism of the opponent in a debate, rather than refutation of the opponent's arguments. This diversionary attack, like the *red-herring* fallacy discussed later, is a smokescreen that uses emotional impact to draw attention away from the relevant logical arguments. Three types of personal attack are:

• *verbal abuse*, in which one directly attacks the opponent's character, personality, or psychological health, although those factors are irrelevant to the argument being 'refuted'.

> "The so-called discoverers of cold fusion are more interested in glory and a Nobel prize than in well-controlled experiments."

• *challenging the objectivity of the opponent*, in which one argues that the opponents' bias forces them to argue as they do, regardless of the argument's validity.

> "It is not surprising that *A* rejects these experimental data, since they refute his hypothesis."

• *'practice what you preach'*, in which one defends oneself by claiming that the opponent is just as guilty.

> "*A* claims that I have ignored conflicting evidence, but she has ignored . . ."

• **Mob appeal** is the assertion that one should accept an argument in order to join the crowd. Mob appeal is the premise for the emotionally enticing conclusion that 'right thinking' people are a group of winners. Different manifestations of the mob appeal fallacy are:

• *mob psychology*, in which the arguer seeks a simultaneous group response through the bait of inclusion or the threat of exclusion. Politicians and preachers use this technique; scientists do not.

• *bandwagon*, in which it is claimed that the group knows best and an individual is in danger of being left out.

> "Everyone's accepting this new theory and finding applications for their own field."

> "In science the authority embodied in the opinion of thousands is not worth a spark of reason in one man." [Galileo Galilei, 1564-1642]

• *egotistic appeal*, which provides a simple way for an individual to be like someone famous.

> "Nobel prizewinner *A* advocates the hypothesis, so wouldn't you?"

• *status symbol*, in which the individual joins a group composed only of the superior people.

> "Mensa is the most fascinating of clubs, because only those with intelligence in the top 2% can join."

> The last football game that I attended was USC versus UCLA, in about 1969. It was called a 'great game': the sides were evenly matched and the team I favored (USC) came from behind to win in the last couple of minutes. My overriding memory, however, is that the fans on both sides were chanting "Kill! Kill! Kill!" and meaning it. They cheered each time a member of the opposing team was injured or carried from the field, and the game was dirty enough that such incidents were fre-

quent. That day I lost my interest in football, but I gained realization of the need to make my own judgments, rather than accepting mob opinion.

• **Might makes right** is the argument that the listener must accept the arguer's conclusion or suffer the consequences. The threat may be physical or it may involve some other undesirable action. Between parents and children it is usually the former, and among scientists it is the latter. The threat is irrelevant to the validity of the conclusion, and yet it may affect the listener's decision-making.

> "Everyone knows that this theory is correct, and if you try to prove otherwise you will destroy your credibility as a scientist."

• **Extenuating circumstances** is the plea for accepting the conclusion out of pity for someone. The arguer claims that acceptance will help someone in trouble, or that rejection of the conclusion will cause undue hardship to someone (usually to the arguer). The extenuating circumstances are irrelevant to the validity of the basic argument.

> Students often use the plea of extenuating circumstances on their teachers; lawyers use it on juries and judges. Scientists use it in personal, not scientific, arguments.

• **Red herring** is diversion of attention from an argument's weakness, by creating a distracting smoke-screen. The fallacy is named after a technique used in training hunting dogs: drag a sack of red herring (a strongly scented fish) across the scent trail that the dog is supposed to follow, and train the dog to stick to the main trail without being distracted or diverted to the red-herring trail. The fallacious *red herring* can and usually does consist of valid reasoning, often protracted and sometimes emotional, so the listener is left with the impression that the conclusion is valid. In fact, the red herring is a related issue that is extraneous to the central argument.

This style of misdirection is the secret of many magicians' tricks. It rarely is employed deliberately in scientific arguments. However, a similar smoke-screen – 'straining at a gnat and swallowing a camel' [Matthew 23:24] -- is sometimes adopted: the arguers demolish a minor criticism of their argument, giving the false impression of careful and objective thoroughness, while obscuring a brief mention of a serious weakness in the argument.

> When the theory of evolution was proposed by Charles Darwin and advocated by Thomas Huxley, falacious refutations were rampant: Evolution is inconsistent with the Bible (appeal to authority); Darwin is a heretic (personal attack) who should be excluded from our community of scientists (mob appeal) and will certainly burn in Hell (might makes right).

* * *

Faulty Link Between Premises and Conclusion

> "The myth of the scientific method keeps the scientific community from recognizing that they must have a humanly developed and enforced professional ethics because there is no impersonal method out there that automatically keeps science the way it ought to be." [Bauer, 1994]

The hardest fallacies to spot are those that lead to a conclusion with which we agree. No scientist would disagree with Bauer's [1994] conclusion that personal ethical responsibility is essential. And yet, his argument is fallacious: we may value the innate checks-and-balances of the scientific method, but no scientist responds by abdicating personal ethics. Between premise and conclusion, the argument has gone astray -- in this case by misrepresenting the relationship between scientist and scientific method. This *straw man* fallacy is one of several ways in which the link between premises and conclusion may be fallacious.

• **Missing the point** is basing a conclusion on an argument in which the premises actually lead to a quite different conclusion. Like *red herring*, much of the argument in *missing the point* is valid but only appears to lead toward the conclusion. *Red herring*, however, is often deliberate whereas *missing the point* is accidental. The fallacy is detectable by deciding what conclusion is actually warranted from the premises, then comparing this valid conclusion to that drawn by the arguer.

> "Hypothesis *A* fails these two tests, and consequently hypothesis *B* is the best explanation."

• **Overinterpreting** is the attempt to claim a firm conclusion although one or more premises is quite uncertain. The fallacy is in asserting a definite, deductive conclusion:

> "Scientists have tried for years to refute this hypothesis and have failed. Thus the hypothesis must be true."

In contrast, the following somewhat similar argument is valid, because it properly considers the evidence as inductive:

> "Many attempts to find N rays have failed, although N rays should be detectable by these tests. Therefore N rays probably do not exist."

• **Begging the question** is an argument in which the logic may be valid, but a dubious premise is either propped up by the conclusion or is ignored entirely. This term is used to describe two different fallacies: ignored dubious premise and circular reasoning.

An *ignored dubious premise*, omitted from but essential to an argument, is a common pitfall. Ignoring a premise is reminiscent of but more extreme than the fallacy of *suppressed evidence*. This fallacy is one of the most serious pitfalls of scientific research, for three reasons. First, everyone is better at noticing characteristics of observed features than at noticing that something is missing. Second, once a premise is overlooked, it will be harder for anyone else to recall. Third, most occurrences of this fallacy could be avoided, if the researcher would just list the premises. Too often, scientists fail to ask themselves what their premises are, or they think about the answers superficially but fail to write them down systematically.

> "You need a different kind of instrument, because the one you have broke down." The dubious premise is that a different type of instrument will not break down.

Circular reasoning is an argument in which the conclusion and premises seem to support each other, but actually they say virtually the same thing in two different ways. The logic is valid if trivial ($A, \therefore A$), yet the repetition lends the illusion of strengthening the conclusion.

> "It is obvious that the instrument is not working reliably because it gives anomalous results. These results must be wrong because the instrument malfunctioned."

• **Equivocation** is use of the same word in subtly different senses; due to ambiguity of word meaning, fallacious logic appears to be structurally valid. Particularly subject to this fallacy are arguments that repeatedly use qualitative descriptions such as large and small, or good and bad:

> "The hypothesis is slightly incorrect, because there is a slight difference between predictions and observations." The first use of 'slight' means 'small but definitely significant,' whereas the second 'slight' may mean 'small and statistically insignificant.'

> Proof that -1 is the largest integer [Spencer, 1983]: Listing all the integers . . . -4, -3, -2, -1, 1, 2, 3, 4, . . . where '. . .' extends to infinity, we see that nothing has been omitted. But we also know that the largest integer (n) is the only integer for which there is no $n+1$; the only such number is -1.

• **Straw man** is a fallacious strategy for refuting an argument: misinterpret it, refute the misinterpreted version, and then conclude that you have refuted the original argument successfully. The term is a takeoff on the concepts of scarecrows and burning in effigy: imagine that you set up a straw man and easily knock it down, claiming that you have knocked down the real man. The term 'straw man' is sometimes used in a different sense than used here; it can be a 'trial balloon', an idea that is proposed knowing that it will be knocked down but expecting that it will be a productive starting point for further discussions.

Frequently, hypothesis refutations are of the following form: "Let's examine the truth of your hypothesis by seeing how well it fits the following example: . . ." If the argument or hypothesis really should apply to the example, then this technique is compelling. The refutation or confirmation is valid, and any refuted argument must be abandoned or modified. With a *straw man*, in contrast, the original hypothesis was not intended to encompass the example, so the argument is fallacious although the entire logic of the analysis is just as valid. Thus one should evaluate the appropriateness of the example *before* applying it, lest the refutation act as a smoke-screen.

* * *

Case-dependent Relationship Between Parts and Whole

> Cholesterol seems to have surpassed sex as the number-one source of guilt in modern America. Much of this cholesterol consciousness stems from the 1985 National Cholesterol Education Program. All Americans were urged to reduce cholesterol in order to avoid heart disease. Surprisingly, however, there was virtually no direct scientific evidence that cholesterol reduction prevents heart disease in either women or in the elderly, although 75% of heart attacks are in people older than 60 years.
>
> The key studies were all on middle-aged men with high cholesterol. These studies conclusively found that: (1) higher cholesterol level is associated with higher risk of heart disease, and (2) giving cholesterol-lowering drugs to high-cholesterol subjects reduced their risk of heart disease. The first finding established a correlation, and the second result demonstrated causality. Generalization of this pattern to middle-aged women and to the elderly of both sexes is plausible, but neither data nor deduction implies it. [Kolata, 1992b].

The conclusion that everyone should reduce cholesterol is a *hasty generalization*, the extrapolation to an entire population from a possibly nonrepresentative sample. The conclusion *may* be correct -- indeed, it has been demonstrated by subsequent experiments to be correct -- but it does not follow compellingly from *these* data. This inductive fallacy and several deductive fallacies go astray

in linking the individual to the general, or parts to the whole, or the universal to the particular. The validity of such arguments is case-dependent: arguments with identical form can be valid or invalid, depending on the specific relationship between parts and whole.

• **False extrapolation to the whole** is the false conclusion that the whole exhibits some characteristic because one or more parts exhibit it. The argument is structurally valid; whether it is correct or not requires careful evaluation of the content, because sometimes extrapolation to the whole is warranted. For example:

> Invalid: "The mistaken belief that technology is applied science . . . implies that any advance in scientific knowledge could be harnessed to useful applications" [Bauer, 1994]. Actually, scientists argue only that *many* scientific advances have valuable practical applications.
>
> Valid: "This prediction of the hypothesis is refuted, and therefore the hypothesis is disproved."
>
> Invalid: "This prediction of the hypothesis is confirmed, and therefore the hypothesis is proved."
>
> Valid: "This premise in the argument is false; thus the argument is false."
>
> Invalid: "Every premise in the argument is true; thus the argument is true." Remember that the truth of a conclusion depends both on the truth of premises and on the validity of the logic.

• **False extrapolation to parts** is the false conclusion that a part exhibits some characteristic because the whole exhibits it. This potential fallacy is the reverse of the previous one. The conclusion may be either correct or incorrect depending on the content:

> Valid: "The argument is correct (valid and true), so every premise must be true."
>
> Invalid: "The argument is incorrect (either invalid or untrue), so every premise must be false."
>
> Valid: "This journal requires peer review for its papers; therefore this article in the journal has been peer reviewed."
>
> Invalid: "That scientific meeting is worth attending, and consequently every talk at the meeting is worth attending."

• **False extrapolation to the individual** is misapplication of a generalization to an individual case. This fallacy is the reverse of *hasty generalization*, and it is somewhat similar to the fallacy of *false extrapolation to parts*. Like these, it may be correct or incorrect depending on the content. The fallacy lies in ignoring evidence that the general rule is inappropriate to this specific case.

> "Publication is an essential part of any research project. Therefore my manuscript should not be refused publication even if the reviews were negative."

• **Hasty generalization** is the inductive extrapolation to an entire population, based on a sample that is nonrepresentative. Often the sample is too small to be representative, but smallness does not necessarily imply the fallacy of *hasty generalization*. A sample of only two or three can be enough if one is dealing with a uniform and representative property: for example, learning not to touch fire. *Hasty generalization* is frequent among non-scientists; it is the origin of many superstitions. A key difference between scientific and popular induction is that the latter usually ignores the need for a representative sample. The consequence is vulnerability to *hasty generalization*.

Hasty generalization is quite similar to the fallacy of *false extrapolation to the whole*. The two differ, however, in the scopes of their conclusions: a general statement about every member of a population (*hasty generalization*) or the collective behavior of a class (*false extrapolation to the whole*).

> Hasty generalization: "Wristwatches with radium dials are safe, so all radium samples are safe."
>
> False extrapolation to the whole: "The half-life of a radium-226 atom is 1622 years; thus brief exposure to radium poses negligible hazard."
>
> Which is this? "I seldom detect the effect, so the effect must be rare."

* * *

H. H. Bauer, a chemist and self-proclaimed expert in STS (science, technology, and society), succeeded in packing a remarkable number of the foregoing fallacy types into a single paragraph:

> "In what sense, then, are the social sciences actually science? They have no unifying paradigm or the intellectual consensus that goes with it. They have not produced distinctive and reliable knowledge that is respected or valued by human society as a whole. Yet those are the very qualities for which the natural sciences are noted and respected; they are the qualities that we associate with something being scientific -- that is, authoritatively trustworthy. The social sciences are simply not, in the accepted meaning of the term, scientific. And that conclusion has been reached by at least a few practicing social scientists -- for example, Ernest Gellner." [Bauer, 1994]

Bauer's argument succumbs to at least seven fallacies:

• *Suppressed evidence*: His assertion that none of the social sciences has a paradigm is incorrect.

• *Suppressed evidence*: To claim that the social sciences have not produced reliable knowledge, one must ignore countless concepts such as supply and demand (economics), stimulus and response (psychology), human impacts of environmental change (geography), and human impacts of racial and gender stereotypes (sociology).

• *False dichotomy*: He assumes that the accumulated knowledge of all of the social sciences can be classified as either having or not having a consensus. In actuality, consensus is a continuum and degree of consensus varies tremendously both within and among fields.

• *Mob appeal*: He claims that 'human society as a whole' determines the reliability of knowledge.

• *Straw-man*: His definition of 'scientific' as 'authoritatively trustworthy' is not merely a weak assumption; it is a deliberate misrepresentation.

• *Appeal to authority*: He seeks validation for his stance by quoting one social scientist.

• *False extrapolation to the whole*: He applies a conclusion based mainly on sociology to all social sciences.

Chapter 5: Experimental Techniques

> "An experiment is a question which science poses to Nature, and a measurement is the recording of Nature's answer." [Planck, 1949]

Experimental design determines whether a research report is read or ignored, whether a result is accepted or rejected, and whether a scientist is judged superior or inferior. Most scientists and many technicians can carry out an experiment successfully. An experiment's value, however, depends not only on outcome but also on the skill with which it is designed. Fortunately, this skill, which develops with experience, also can be fostered deliberately. This chapter provides a variety of experimental tips, culled from the experiences of many scientists.

Like military planning, research planning has three levels [Beveridge, 1955]:

[Larson, 1985]

- **tactics**, the small-scale and relatively short-term planning of an individual experiment. The key question is 'how'. Tactics must be considered in the context of
- **strategy**, the broader approach to the research problem, which involves an extensive suite of experiments. A strategy is most efficient if it is considered in the context of
- **policy**, the determination made by funding agencies and committees concerning which general problems are the most crucial in a science.

Like business planning, research planning should involve the following:

- **risk analysis.** What is the chance of success? What could go wrong and what would its impact be?
- **risk management**. How can I improve the chances of success? How can I avoid possible factors that would make the final result ambiguous or misleading?
- **time analysis**. How much time will it take for each step? How will potential problems affect this estimate?
- **time management**. How much control do I have over the amount of time required for each step? Where can I streamline the procedure without weakening the integrity of the experiment?

An intelligent person would never go to war or start a business without careful analysis of the factors above, yet most research planning gives them only brief attention. Usually we are so immersed in the details that we neglect the broader concerns. We may even claim that we are so busy doing that we don't have time for esoteric planning. *Careful planning of an experiment determines its value*. If most experiments were to begin as a 'gedanken' experiment, a thoughtful anticipation of the likely progress and outcome of the experiment, then the few that are physically undertaken would be more likely to be key experiments.

> "From the way a war is planned, one can forecast its outcome. Careful planning will lead to success and careless planning to defeat. How much more certain is defeat if there is no planning at all!" [Sun Tzu, ~500 B.C.]

Failure to consider the factors above creates some of the more common experimental **pitfalls**:

• **underestimating the amount of time that an experiment will take**. Underestimation is most acute when the planned experiment has never been done before (e.g., when one is designing new equipment for a novel experiment). Almost always, one's overall time estimate is much shorter and less realistic than would be an estimate based on a list of the time requirements of individual steps. Most experimenters also fail to include time estimates for delays, setbacks, and unexpected problems, and their time estimates assume 'production mode' rather than the entire period from set-up to shut down. I routinely allow optimism to carry me into this pitfall, even though I recognize the wisdom of my wife's rule of thumb: *carefully estimate the time for each individual step, sum these times, and then double the total*.

• **lack of time management**, resulting from taking each step as it is encountered. For example, running batch processes is usually more efficient than doing experiments in series. The wait time that occurs somewhere in most experiments can often be filled with another part of the experiment, if the process is planned as a whole. If I concentrate, I can keep three processes going at a time.

• **lack of strategy**. Even a tactically brilliant project can be strategically ineffectual or foolish; the best example is the Japanese attack on Pearl Harbor. The consequences of clever implementation of unfocussed research are less drastic: inefficiency and ho-hum science. Many ingenious experiments contribute little, because of insufficient attention to overall strategy. Too many experiments are selected merely because they are obvious or logical follow-ups to previous experiments. A more powerful selection strategy is to consider various possible experiments, then select the one that is likely to contribute most.

• **lack of risk management**. Often, surprisingly small changes in design or technique can profoundly affect the value of an experiment. Yet these refinements are neglected, because planning is short-circuited by optimism, lack of risk analysis, or enthusiasm to get started. In hindsight, the changes that should have been made are obvious.

> "The winner does everything to ensure success before he fights. The loser rushes into combat without adequate preparation." [Sun Tzu, ~500 B.C.]

* * *

Observational versus Experimental Science

Many scientific disciplines are more observational than experimental. Within these research areas, only a few of the guidelines for experimental design in this chapter will apply. For example, in observational or descriptive branches of biology, ecology, psychology, anthropology, and astronomy, manipulation of variables is not always possible. With many natural phenomena one cannot

control experimental conditions. Yet the basic elements of scientific method are identical to those used with other experiments: observations inspire hypotheses, which can be tested only with further observation.

Scientists generally use the term 'observations' as a synonym for 'data', whether or not the experiment actively manipulates the observed environment. This section, in contrast, focuses on that distinction. Unlike experimental science, much observational science is akin to the Chinese concept of *wu-wei*, or 'not doing'. Wu-wei is a balance of active and receptive, an alertness coupled with a willingness to allow nature to unfold and reveal itself.

Throughout scientific history, some scientists have chosen this method of alert receptivity. Greek science was almost entirely observational; it sought order in the universe through observation, interpretation, and classification rather than through experimentation. Charles Darwin, as biological taxonomer on the Beagle, recognized patterns that allowed him to understand both the evolution of coral reefs and the process of biological evolution through natural selection. Within the science of geology, a more observational science than most, there are highly experimental fields such as experimental geochemistry as well as largely observational fields such as paleontology. The lack of experimentation in paleontology has not prevented the field from providing the age-dating foundations for most of the rest of geology, or from revealing a wealth of climatic and evolutionary information.

Observation is the primary method for understanding complex systems. Control of all possibly relevant variables in such systems may be impossible, and the theories may be so simplified that they cannot predict the observations reliably. In studying complex systems, the search for one phenomenon frequently reveals an even more interesting phenomenon.

The approach to observational science often begins qualitatively, as a search for an order that characterizes the system. Usually the researcher observes many variables, hoping to detect any patterns. These patterns or relationships may or may not be causal. If an apparent pattern is found, a suite of later observations can be designed to test its significance.

An observational science can evolve into a more experimental science, particularly when a new paradigm guides observations by imposing order on the complexity and indicating which parameters are important for study. Astronomy is a good example: for thousands of years it was purely observational, then it became capable of quantitative predictions such as the occurrence of seasons and eclipses. Since Newton, work in astronomy has been increasingly quantitative and theoretical. Even without the ability to affect the planets and stars, astronomical experiments can isolate variables and test hypotheses.

Unlike hypothesis-driven experimentation, with its limited choice of expected results, observational science often yields unpredicted results. While this can be advantageous, promoting insight and creativity, there also are drawbacks. Unexpected results often are overlooked or rationalized (see Chapter 6). A particularly challenging aspect of observation is the necessity of noticing absence of a phenomenon; absence can be as significant as presence. For example, consider Sherlock Holmes's search for the perpetrator of a break-in:

> " 'Is there any point to which you would wish to draw my attention?' [asked Inspector Ross].
>
> 'To the curious incident of the [watch]dog in the night-time,' [answered Sherlock Holmes].
>
> 'The dog did nothing in the night-time.'
>
> 'That was the curious incident,' remarked Sherlock Holmes."
>
> [Doyle, 1893a]

Observational science has a bad reputation among some scientists, for several reasons. First, it cannot change experimental conditions as some other fields can. Second, its initial stage is often 'just' data gathering, a 'fishing expedition', because the phenomenon is still so incompletely understood that few hypotheses are available to channel observations. Third, and probably most important, the initial stage of most observational sciences is qualitative -- subjective -- not quantitative. Often the system is so complex and so many parameters can be measured quantitatively, that the scientist cannot discern which characteristics should be measured. I suspect that astronomy enjoys a much higher reputation than clinical psychology among scientists because it has progressed farther along a continuum, which begins with overwhelming complexity, progresses to pattern recognition and qualitative hypothesis testing, and culminates in quantitative testing of theoretical models.

Some scientists whose research is amenable to carefully designed experiments think that any research lacking such control is 'less scientific' than their own research. Nothing could be farther from the truth, and one should beware the assumption that the same standards for scientific method apply to all types of science. "If one has only a hammer, one tends to look at everything as if it were a nail."

> "More discoveries have arisen from intense observation of very limited material than from statistics applied to large groups. The value of the latter lies mainly in testing hypotheses arising from the former." [Beveridge, 1955]

The following late 19th-century experiment by naturalist J. Henri Fabre [Teale, 1949] is worth a detailed description because it illustrates both the power of observational techniques and the remarkable impact that the subtlest experimental intervention can have on the value of a suite of observations. It should be noted that Fabre tried several similar, unsuccessful experiments before he recognized an emerging opportunity and seized it.

> To what extremes can animals be enslaved by instinct? Fabre investigated this question by studying the pine processionary, a moth caterpillar that leaves a silky thread behind it as it travels. Nearly always, it chooses to follow existing silky paths, laid down by other pine processionaries. Usually this strategy is valuable for survival, leading each individual among the established food supplies of pine needles.
>
> Watching a parade of pine processionaries approach the rim of a palm vase in his greenhouse, Fabre waited until the leader had completed a full circle around the rim. He intercepted the parade, by quickly brushing away all caterpillars and trails below the rim. That was the extent of his experimental intervention. He then observed the result, a continuous, leaderless string of caterpillars rimming the pot.
>
> Because pine processionaries simply follow the silky trails left by others, the caterpillars kept going in a circle. Night and day, they kept going in the same circle. Only during the coldest part of the night did they slump. When they started again on the third day, they were huddled in two groups. Two leaders started the march around the rim, but soon the two groups combined into a continuous ring. On the fourth day the first to wake had slumped off the track. It and six followers entered the new territory of the pot's interior, but they found no food and eventually wandered back to the rim, retaking the circular path. On the fifth day, a leader and four followers strayed from the path and explored the outside of the vase, to within nine inches of a pile of pine needles. They failed to notice this food and wandered back to the rim. Two days later, now staggering from hunger, one wandered and found the pine needles. Eventually the rest of the group followed. The pine processionaries had circled hundreds of times over a seven-day period and failed to recognize the uselessness of their circular path. They followed instinct to the point of collapse repeatedly, surviving only by chance.

* * *

Seizing an Opportunity

Fabre's experiment above is a classic example of the power of seizing a scientific opportunity. Often scientists undertake carefully planned experiments, but occasionally chance presents them with an opportunity. Such opportunities are most common in the more observational fields of research.

For example, whenever an earthquake happens, teams of seismologists rush to the scene with their seismometers, in order to monitor the aftershocks. When the 1971 San Fernando earthquake struck, I was at Scripps Institution of Oceanography, less than 100 miles away. I failed to seize the opportunity: I slept through the earthquake. My advisor, in contrast, ran into the bathroom and looked into the toilet. He saw that the water was sloshing north-south. Because he knew that active faults lie north but not south of San Diego, he inferred that a major earthquake had just struck north of us -- in the Los Angeles region.

A few hours later, he and several other geologists (myself included) were driving near San Fernando, looking for the fresh fault scarp. At least that is what we were trying to do; actually we were stuck in a canyon in what seemed to be the world's largest traffic jam, while earthquake-loosened pebbles bounced down the hillsides and pelted the cars below. I remember wondering whether seizing this opportunity might be the dumbest idea I had ever gone along with.

Seizing an opportunity has sometimes been used as an excuse for skimming the cream and discarding the milk. In Egyptology, for example, the early approach was to grab the spectacular, expending no time for 'details' such as careful documentation of the less glamorous debris or post-excavation restoration of the site environment. Now archaeological work in Egypt is more careful throughout each project [Cowell, 1992], because an archaeological site offers no opportunity for a second chance or replicate study.

> Supernova SN1987A was a successful example of scientists seizing an opportunity [Browne, 1992]. This explosion of a star occurred so close to earth ('only' 160,000 light-years away) that it was visible to the naked eye. It was the closest supernova in the last 400 years, an astounding chance to exploit modern astronomical technology to verify popular but untested models such as that of neutrino flux. The challenge was that study of SN1987A required a very fast scientific response, because the supernova peaked in brightness only three months after it was discovered. Both astronomers and funding agencies bypassed existing plans and procedures, achieving a sudden burst of observation, confirmation, and modification of theories.

* * *

Experimental Equipment

Equipment, not rare opportunities, is the mainstay of most experimental science. The applications, complexities, and costs of research apparatuses differ, yet several concerns and potential pitfalls are common to most equipment used in science.

Invention often follows technology. When new technology permits higher-resolution studies or a novel type of measurement, new perspectives often result. One should be alert for such techno-

logical developments, even if they arise from outside one's specialty, because of the potential for cross-disciplinary applications.

New technology also has potentially serious pitfalls. First, fascination with the new and complex can prevent objective evaluation of a new device's strengths and weaknesses. For example, as I write this, the most powerful of the supercomputers is the Cray. Many scientists are impressed with results from the Cray. Some claim that anything produced on it must be right, and that its calculations supersede those from other computers. In fact, all computer calculations are subject to the same pitfalls of programming error, rounding error, and invalid assumptions; the supercomputers merely allow faster completion of complex calculations.

Researchers often are faced with a choice between two pieces of equipment: an older and a newer model. Perhaps one already has the older type and is thinking of acquiring the newer version. Usually the newer design uses state-of-the-art technology and therefore is more expensive, more efficient, and more accurate. Will enough experiments be undertaken for the greater efficiency to justify the greater cost? Cost of experimenter time must be weighed against equipment cost. Similarly, one must weigh the option of obtaining more measurements with lower accuracy against that of fewer measurements with greater accuracy. The latter is more aesthetically pleasing but not necessarily the most practical solution, and simple statistical analyses can help in this comparison.

Occasionally investigators choose to design their own apparatus, perhaps because none is commercially available or because personally constructed equipment is more suitable or less expensive than commercial. Almost always, this design and construction takes more time than expected. Yet home-built equipment also has several advantages, such as intimate familiarity by the researcher. Wilson [1952] gives a detailed review of factors to consider when designing and building one's own equipment.

Whether using old or new equipment, the most frequent equipment pitfall is trusting the equipment. *Nearly all equipment needs standards and calibration*, regardless of what the manufacturer may imply. The need for calibration is obvious with home-built equipment, but calibration checks are just as necessary for sophisticated, expensive equipment. Indeed, this pitfall is even more insidious with the newer, higher-technology equipment. Digital displays and direct computer interfacing of equipment do not assure reliability.

Precision and accuracy, once determined, cannot be assumed to persist unchanged. Both can be destroyed by equipment malfunction and by subtle changes in the experimental environment. For example, I once subcontracted to another lab for 400 chemical analyses. In examining the data and the replicate measurements of standards, I found that the final 25% of the analyses were worthless. A power cord had been replaced and the equipment was not recalibrated after this 'minor' change.

Creating or purchasing some standards, then occasionally running them to confirm equipment performance, takes trivial time compared to the span of routine measurements. In contrast, lack of calibration checks can mean that entire experiments have to be redone. If realization of data unreliability dawns after publication, the setback can affect an entire research discipline.

* * *

Prototypes and Pilot Studies

When designing a new apparatus for a suite of experiments, it is usually a good idea to build a prototype first. When beginning a novel type of experiment, it is usually a good idea to do a pilot study first. In both cases, it is tempting to skip this step 'to increase efficiency'. Skipping this step

is almost always a false economy, unless the new apparatus or experiment is only a slight refinement of a familiar one.

The experimental prototype is a routine step in applied science, where it provides a bridge between theory and routine practical application. Applied science usually has two scales of prototype: laboratory prototype and then pilot plant. Only after both have been tried does a company decide whether commercial production is viable.

The prototype is a less common step in basic research, although some of the same factors that encourage its use in applied science apply to basic science. The prototype allows feasibility study, detection of practical problems, and improvement of design parameters. It also permits identification and isolation of unanticipated factors that could affect the success or accuracy of the experiments. Consequently, far different standards and objectives apply to the prototype than to the final apparatus:

• The prototype is much less expensive and time-consuming to build than the final apparatus. Often it is smaller, less robust, and less streamlined.

• The prototype is much more versatile than the final apparatus. Modification is easier, parts can be exchanged, and 'quick and dirty' fixes are acceptable.

• Depending on the type of measurement, the prototype may have a smaller or larger range of measurement values than the final apparatus will need to have.

• Precision and accuracy may be lower on the prototype, and both are improved as problem variables are isolated. The prototype is not necessarily well calibrated, because we are probably more interested in sensitivity analysis than in accuracy of these results.

• Measurements may be more cumbersome and slower on the prototype than on the final apparatus.

A prototype is likely to be needed whenever equipment design is substantially modified. It can even be a worthwhile time saver when one is building equipment or an experimental setup similar to published ones and depending on rather scanty published details. It is better to discover that the author left out a 'slight complication' when trying out a prototype than in the midst of an experiment.

The pilot study is the procedural analogue to an equipment prototype, and many of the considerations above apply equally well to pilot studies. Different standards [Beveridge, 1955] concerning variables and their control apply to pilot studies than to the formal experimental series:

• One can use extreme values for a variable in the pilot study to see if they have any effect. If they do seem to have an effect, then the formal experiment can focus on the range of most interest for this variable. At that stage, higher accuracy and precision probably will be required. Indeed, statistical analysis of the pilot data can indicate how many measurements will be needed to detect the effect (Chapter 2).

• In some situations, many variables could have a significant effect, and it is not obvious which needs to be incorporated into design of the formal experiment. One can lump many variables in the pilot study. Of course, some caution is needed to prevent cancellation of the effect of one variable by the opposing effect of another. This approach is most effective if one knows the probable direction of the potential influence of each variable. If a composite effect is found, then formal experiments can be designed that will systematically isolate the effects of each variable.

Pilot studies can indicate whether a potential experiment is likely to be fruitful and how one should deal with the relevant variables. Pilot studies cannot substitute for a well designed formal experiment.

Prototypes and pilot studies are modes of scientific troubleshooting. Whether or not we expect problems, these techniques help us to avoid them.

* * *

Troubleshooting and Search Procedures

Troubleshooting is a familiar, intimate part of science. The trouble may involve computer hardware or software, malfunctioning equipment, or an experiment that is giving results that are unexpected and possibly unreliable. These and many other problems are solvable with established troubleshooting and search procedures. Yet the techniques are published in few places, and most of us react to encountered problems by thinking of only one or two remedies. Wilson [1952] considers troubleshooting and search techniques in detail, and the completeness of the following discussion owes much to his comprehensive treatment.

The foremost rule of troubleshooting and search is: *keep records* to avoid duplication of effort and floundering, to reveal any patterns in the troubleshooting results, and to make it easier to identify potential tests that you have overlooked. Keeping records is unnecessary for the first or even second attempted solution. As soon as troubleshooting extends beyond a few minutes, however, one should start jotting down notes of what has been tried and what it yielded.

The frustration implicit in troubleshooting can result in needless damage. Hippocrates was familiar with the problem 2000 years ago. His guideline, which could have supplanted later leechcraft, is still apropos: *Primum non nocere*; first do no harm. When diagnosing a medical problem, exploratory surgery is an acceptable last resort; autopsy is not.

A subtler manifestation of *primum non nocere* is the following question: is the object of the quest worth the cost of the search? Cost can take various forms, tangible and intangible. When cost is computed in dollars, this question is the daily quandary faced by NSF and the occasional topic of intense arguments, as exemplified by the debate over star-wars research.

If troubleshooting new equipment:

1) Remember the facetious saying, "If all else fails, read the manual." Probably something is connected wrong, a setting is incorrect, or a step is being left out. The better manuals even have a section on troubleshooting.

2) If possible, run a standard that you know is supposed to work on this equipment. If the standard works OK, then how does your sample differ from the standard? If the standard doesn't work either, then go on to the next step.

3) Examine all of the equipment for visible signs of damage.

4) Try to isolate which part of the equipment is malfunctioning. Some of the search procedures discussed later may help. Sometimes it is possible to swap out parts of the equipment. Some parts can be tried in isolation or in conjunction with other working equipment, and some circuits can be tested with a multitester (AC, DC, and resistance).

5) Scan the following sections for any hints that might be relevant.

6) Call the supplier or manufacturer, tell them that *we* have a problem with the new equipment, and try to get troubleshooting help over the phone. Why not do this first? Telephone help is fine, but if the call merely results in arrangements for a replacement requiring delays of days or even weeks, then a few minutes or even hours of preliminary troubleshooting may be justified. The supplier may suggest that returning the equipment for their evaluation is more practical than telephone troubleshooting, because that is easier for them. If so, remind the supplier that they claimed to be in business to serve you, not vice versa.

If troubleshooting equipment or an experiment that formerly worked fine:

1) Go back to previous data and identify when the problem began, then list all changes to the system that occurred at about that time. The cause of the problem is probably in the list.

2) Run a benchmark check: try to replicate a measurement or result that you have previously obtained and that you are reasonably certain is valid. If it replicates OK, then how does this sample differ from the problem sample? If it does not replicate, then what may have changed since the original correct measurement? If this test is inconclusive, then a second replication test may be worthwhile, using a sample with quite different characteristics.

3) Consider the following frequent sources of equipment problems: incorrectly remembered measurement procedures, blown fuse or circuit breaker, part failure, a corroded connection, supply voltage variations, and temperature-sensitive components or equipment response. The first three usually cause a complete failure, and the others often cause intermittent problems.

4) If none of the above help, then possibly you have an uncontrolled variable that is influencing the results. Methods for dealing with such variables are described later in this chapter.

* * *

Search is fundamental to scientific method. Search procedures can be used for finding objects and for troubleshooting problems. More generally, search is exploration-based research. Search procedures can provide a practical way of dealing with the complexity of nature. They can help one to focus efforts or scientific questions, in order to reduce them to a tractable size (Killeffer, 1969).

Most scientists are aware of most search procedures. Nevertheless, we often succumb to the pitfall of choosing the first search procedure that comes to mind, rather than deliberately selecting the most appropriate procedure. The following list of **search considerations and techniques** is largely based on a compilation by Wilson [1952]:

• **Characterize the object of the search**. List the characteristics of the search object, and for each characteristic consider whether or not the object differs from its surroundings.

> "Kilimanjaro is a snow covered mountain 19,710 feet high, and is said to be the highest mountain in Africa. Its western summit is called by the Masai 'Ngaje Ngai,' the House of God. Close to the western summit there is the dried and frozen carcass of a leopard. No one has explained what the leopard was seeking at that altitude." [Hemingway, 1940]

> Characterizing the search object has been the nemesis of attempts to find the 'missing' mass of the universe. If, as many cosmologists previously expected, the universe is to collapse someday into an infinitely small point like that which began the

big bang, then there must be enough mass in the universe to produce the required gravitational attraction. Yet only about 10% of this mass has been found to-date [Wilford, 1992c], even including both what has been observed and what has been extrapolated to exist. Imagine the challenge of searching for something when you don't know what it is and you don't know where it is. Remarkably, astronomers are finding at least some of this dark matter, by observing galaxies whose motions require gravitational forces far larger than the observed masses could generate.

• **Pick the most efficient method of detection**. For example, select a technique that sees the search object but not the surroundings. In picking the most efficient method, consider the effort, time, and money needed to acquire, set up, and employ each method.

Diamond hunter Gavin Lamont discovered a gigantic diamond-bearing kimberlite pipe in Botswana, although the entire exploration region was covered by 160 feet of surface sediments that contained no diamonds. He used one fact, one speculation, and months of perseverance. The fact was that even though diamonds are incredibly rare, they are accompanied by common indicator minerals garnet and ilmenite. Thus one should search for garnet and ilmenite first. His speculation was that the only way garnet and ilmenite could get to the ground surface from 160 feet down would be reworking by deeply burrowing termites. Therefore he ran a search pattern through hundreds of square miles, easily finding each termite mound sticking up from the flatlands, and examining the termite mound for the presence of the dark minerals garnet and ilmenite. When he finally found the indicator minerals, he sank a shaft to what later became a 4.5-million-carat-a-year diamond mine [Heminway, 1983].

• Before embarking on a major search, try to **establish that the object really does exist in the area being searched**. For example, do not spend a major effort taking equipment apart if the problem could be elsewhere (e.g., power fluctuations).

A friend of the Mulla Nasrudin found the Mulla crawling around outside at night beneath a lamp post. Of course he asked, "What are you doing?" "Looking for my key", replied the Mulla. The friend asked, "Where did you lose it?" and the Mulla replied "In my house". The exasperated friend asked, "Then why are you looking for it here?" The Mulla answered, "Because there is more light here." [Sufi teaching story, e.g., Shah, 1972]

• **Confirm that you would detect the object if you encountered it**. For example, it might be feasible to use an artificial substitute and see if you detect it. Test your detection method at intervals to be sure that it is still working and that it has sufficient sensitivity. Conversely, find out if your test is so sensitive that it gives false positives, i.e. it claims that you have found the object when you have not. False alarms may be acceptable, if you have another test for the object that you can apply at each apparent detection.

On my last oceanographic expedition, we heard an introductory lecture on drilling technology, complete with a 70-pound, 1'-diameter drill bit as a visual aid. After the lecture, two women scientists saw that the drill bit had been forgotten. "Let's put it in the men's room and stick a toilet brush in it," they decided, "No man will ever recognize it." For more than a week, they were right.

• **Keep a record of the search**; ideally, **flag searched areas** so that they are readily distinguishable from unsearched regions.

• **Search the most probable place first**, then search places that are successively less likely to contain the object. Use specific criteria for estimating promising locations; do not just play hunches.

> Meteorites are occasionally found, but until recently they were considered to be far too rare to search for. Now they are being found in record numbers in Antarctica, and scientists have even found eight that they think have come from Mars and several that definitely came from the moon [Gleick, 1992b]. The new success is due to letting natural processes concentrate the meteorites: they are found on top of the Antarctic ice, in environments that are undergoing ablation rather than precipitation.

• **Search systematically**. Haphazard searching can take just as long as a systematic search, yet search some areas several times and others not at all. At the end of a haphazard search one still cannot exclude the searched area. Most searches begin haphazardly, but once they become time-consuming, one should pause and plan a systematic search. Systematic searches are not necessarily inflexible, inefficient searches.

> Paul Ehrlich's hypothesis was that some substances are able to kill parasites without damaging the host. He systematically investigated 605 compounds without success; the 606th, salvarsan, proved effective against syphilis. [Beveridge, 1955]

• **Distribute** your available searching **resources** -- whether time or manpower -- **appropriately** in the different regions. For example, if several areas seem equally promising but some are much easier to search, search the easy ones first. If you will have to quit searching after a limited time, usually a detailed search of the most promising area is more effective than a widespread superficial search. If a little-known phenomenon or a totally new type of observation is being explored, the initial search should probably be a broad reconnaissance rather than a detailed examination of a small subset. Detailed focus is higher risk, until the reconnaissance establishes which parts of the whole are most likely to reward close-up studies.

• **Use a convergent search procedure**, if possible. Convergent searches employ feedback on whether they are getting closer to or farther from the object. This technique is feasible for questions such as "When did the equipment start giving strange results?" or "Where is the short circuit?" but useless for questions like "Where is the needle in the haystack?" When using a convergent search, it is better to overshoot than to undershoot; this is the tactic used in golf putting and in weighing (use large weights until overshooting, then smaller weights). The ideal search procedure eliminates half the possibilities at each step:

> Consider the game of twenty questions, as employed in the old television show 'What's My Line?' There are thousands of professions, yet the questioners were often able to guess the contestant's profession. If one can design each yes/no question to cut the number of possible solutions in half, then twenty questions can sort out one choice from 1,048,576 possibilities (2^{20}). Twenty guesses or a million? Clearly a systematic search procedure such as this one can be extremely powerful. Unfortunately,

most search problems cannot be cast in this form. Furthermore, the technique fails if one of the answers may be incorrect.

• **Use a search method that indicates both direction and distance** to the object at each search step. On the few occasions that this technique is feasible, it is extremely efficient.

An old parlor game begins with everyone except the subject selecting some object in the room. Then the subject attempts the seemingly impossible task of identifying that object. As he wanders around the room, he detects people's involuntary reactions to his proximity to the target object.

• Consider the **probability of missing the object** even if it is in the search path. Decide whether it is more efficient to do a detailed search first or to do a quick reconnaissance first and, if unsuccessful, then do a slower search of the same area. Do not eliminate any area unless there is virtually no chance of having missed the object.

• Consider possible **impacts of the search itself** both on the object and on the difficulty of finding the object. For example, will the search technique preclude re-search? If so, we must be certain that the initial search does not overlook its objective.

Surprisingly often, the following technique helps to solve equipment problems: take the equipment apart and put it back together. Perhaps this procedure will reveal the problem (e.g., a corroded connection) or perhaps it will solve the problem without revealing it (e.g., a loose connection). This procedure is counterproductive, however, if one puts the equipment back together incorrectly and creates another problem; it is much harder to troubleshoot two independent problems than to identify one.

Searching for submarines requires one to consider the impacts of the search: the submarine may detect you before you detect it, and it will take evasive action. Naturalists have the same problem with finding wildlife. Hunters sometimes take advantage of this phenomenon by using beaters to flush the game.

• For multidimensional search problems, **search one dimension at a time**. For example, when searching an area of ground, run parallel lines, with minimal but finite overlap of swaths. For multiple independent variables, check one variable at a time. For example, when tuning several controls for maximum sensitivity, maximize response with each separately, then repeat the procedure if interactions are possible.

• Consider the possibility of the object being present or visible only intermittently. This possibility applies not to physical objects but to a problem or searched-for phenomenon. **Intermittent phenomena** usually require long-term monitoring rather than, or in addition to, a more conventional searching of different areas at different times.

• Consider the possibility that **two independent factors** are helping to hide the object. Would the search procedure be successful regardless of the relative importance of these two controls, or does it

only help to identify one of them? These two-variable problems, like intermittent phenomena, can create baffling, frustrating search and equipment-troubleshooting puzzles. Somehow one needs to separate the two (possibly unknown) factors, as in the one-dimensional search procedure above.

When the search succeeds, **minimize recurrence problems**:

• prevent another loss of the object, or

• make the solution a permanent one rather than a temporary fix-up, or

• make a permanent record of the solution or successful search technique, and assure that this record does not have to be searched for. For equipment problems, paste a label on the equipment describing the symptom and solution.

* * *

Each of the search tactics above can be invaluable on occasion. For particularly intractable search or exploration problems, it can be worthwhile to scan the techniques above, considering the appropriateness of each. Many search problems, however, can be solved by a relatively simple series of steps:

1) describe the problem thoroughly;

2) list the differences between 'signal' and 'noise';

3) evaluate the potential benefit of each difference for the current problem;

4) consider employing a discrimination of more than one difference, either in series (first cut then final discrimination) or in parallel (simultaneous).

* * *

Problem: Find a Needle in a Haystack.

Hint #1: First, define the problem more thoroughly. Ask, "Why do you need to find it?" This is not just a smart-ass question. It is a recognition that there may be many ways to solve a problem, and individual requirements determine the optimum approach. Are you after the needle or a needleless haystack? Which is dispensable: the needle, haystack, or both? Is this a one-time or repetitive problem? Are you certain that the haystack contains only one needle? How critical is it that no mistakes are made?

What is the best way to find a needle in a haystack, given each of the following scenarios?

• Feeding this hay to your thoroughbred horse could give it a punctured stomach.

• The only possible supplier of hay for your thoroughbred-horse stable provides hay that sometimes is unavoidably contaminated with needles.

• You are doing some sewing, and you just lost your needle in the haystack.

• A valuable golden needle is lost in the haystack.

Hint #2: Before deciding on a technique, list the characteristics in which the search object can be distinguished from background:

needle	hay

metal	vegetable
silver-colored	tan-colored
not flammable	flammable
magnetic	nonmagnetic
denser than water	floats on water
sharp	dull
rigid	flexible
etc.	etc.

This listing reveals multiple possibilities that can be considered with the question of hint #1, to select the optimum strategy. A pitfall, however, is the possibility that the needle may be stuck into a straw and thereby exhibit straw-like properties.

Answers (not in order):

- Burn down the haystack, then sift through the ashes.
- Buy an airport X-ray machine and pass all hay through it.
- Throw away the haystack and buy another (hay is cheaper than time).
- Go get another needle.

* * *

Problem: Search for the Top Quark.

High-energy physicists, needing to test theoretical predictions concerning subatomic processes, attempted to detect a subatomic particle that they called the top quark. They wrote computer programs to scan through a very large number of potential events and discard most of them. The remaining events then could be examined more carefully, to see if they might be caused by the top quark. Effectively, hay was plentiful, and the needle was only hypothesized to exist.

A major concern in designing the computer program was the relative impact on the experiment of two types of errors. An alpha error is a false positive (calling a straw a needle). A beta error is a false negative (missing a needle).

As I write this, the top quark finally has been detected.

* * *

Tips on Experimental Design and Execution

> "The general who understands the advantages of varying his tactics really knows the art of war.
>
> "The general who does not appreciate the need to vary his tactics cannot turn natural advantages to account. . .
>
> "The wise man considers both favourable and unfavourable factors, the former to pursue his objectives and the latter to extricate himself from difficulties." [Sun Tzu, ~500 B.C.]

> "In almost every game of chess there comes a crisis that must be recognized. In one way or another a player risks something -- if he knows what he's doing, we call it a 'calculated risk'.
>
> "If you understand the nature of this crisis; if you perceive how you've committed yourself to a certain line of play; if you can foresee you've committed yourself to a certain line of play; if you can foresee the nature of your coming task and its accompanying difficulties, all's well. But if this awareness is absent, then the game will be lost for you, and fighting back will do no good." [Reinfeld, 1959]
>
> "Genius . . . means transcendent capacity of taking trouble." [Carlyle, 1795-1881]

Preparation, experimental design, experiment execution, data analysis, and interpretation are all essential aspects of most research projects. Earlier sections of this chapter discussed experimental design with minimal reference to these companion facets of research, but here we will consider experimental design in the context of the overall experiment. Drawing on useful summaries by Wilson [1952], Killeffer [1969], and Open University [1970], I list the main initial steps in a research project (from conception through experiment), along with tips and guidelines on successful execution of each step:

1) state the general problem.

- What is the objective? Focus on a specific hypothesis; don't undertake a fishing expedition.
- Is the experiment necessary? Can the question that the experiment hopes to address be answered by simply evaluating the hypothesis and its implications critically? Can the problem be solved by finding relevant data that are already published? Is it merely a puzzle that interests you (a perfectly valid reason) or does it affect interpretation of other problems? If the latter, is it a minor or major factor?
- Can the problem be restated in a form that makes it more feasible to solve? Should one test a simplified perspective or include refinements? Does the problem need to be broken down into components that are tested individually?
- What assumptions are implicit in the experimental design? Could the outcome of the hypothesis test be affected by an invalid assumption?
- What is the crux of the problem, the critical unknown aspect? What is the crucial, decisive experiment? Don't settle for one that is merely incrementally useful, but don't reject all potential experiments as not decisive enough.

2) thoroughly **review existing data** on the research topic. Review evidence on the more general problem, to the extent that time permits.

- Methodically summarize assumptions, data, interpretations, and speculations of previous relevant studies. Include your evaluation of reliability and possible weaknesses of each.
- Identify the critical deficiency of previous work.

3) select the most promising experiment.

- Seek a compromise or reconciliation of critical needs with viable techniques.

4) decide how to deal with all relevant variables.

- List all known variables that might influence the result. Classify each variable as either: (a) controllable, (b) uncontrollable but with an approximately known value or known influence on the key dependent variable, or (c) uncontrollable and unknown in influence.
- Decide which variables are of greatest interest. Try to minimize effects of all other variables (e.g., by keeping them constant or by randomization).
- Select one of these tactics: (1) focus on only one variable and vary it systematically; (2) analyze several variables simultaneously through use of a factorial design; or (3) analyze several variables sequentially in a series of experiments.

A later section of this chapter, 'Control of Variables', discusses these options in more detail.

5) choose the equipment to be used, if any.

- Consider the relative advantages of buying, borrowing, and building equipment. Preparations and lead time are greatest for building, less for buying, and least for borrowing. Ability to tailor the equipment to your needs is greatest for building, less for buying, and least for borrowing. Costs are high for both building and buying compared to borrowing. Borrowing is OK for a few experiments but usually impractical for a protracted suite of experiments. Experiments on borrowed equipment tend to be done in binges, with less opportunity for intermediate analyses during experiments and for follow-up experiments.
- Before using equipment, learn its background theory, operations steps, operational considerations, and potential problems. Obviously, some compromise is needed between the ideal of comprehensive understanding and the reality of time constraints. Generating unreliable data and then troubleshooting can be much more time-consuming than learning how to operate the equipment properly and within its limitations. One need not become an electronics technician to use electronic equipment, but pitfalls abound for those who use equipment that they understand only minimally. For example, the dilettante may omit implied operations steps, use the equipment outside its design range, overlook variables that affect equipment results, and misinterpret results.

6) calibrate equipment, both before and during the experiment.

- Test the equipment before starting the experimental series. Do not assume that it can be trusted simply because someone else recently used it successfully. Their samples may have been subtly different, or the equipment response may have changed.
- Choose standards and a calibration procedure appropriate for the equipment, samples, and anticipated data range.
- Recalibrate after the equipment is repaired, moved, or changed, and after any substantial hiatus.
- Run calibration samples regularly, preferably in a randomized mixture with the experimental samples. If equipment response changes with time or warm-up, calibrating at the start or end of each day is insufficient.
- Run blanks if feasible.

7) include **replicate measurements** in your design, if possible. Normally it is unnecessary to replicate every measurement. Replicating perhaps 5% of measurements, on randomly chosen samples and standards, gives a good estimate of overall precision.

8) in the experimental design, **never change more than one variable** or experimental aspect **at the same time** (unless you are using a factorial design).

9) list the entire planned experimental procedure. Do not simply try to visualize the entire procedure in your head.

- Calculate how many measurements or samples will be needed.
- Visualize every step, imagine what could go wrong, and seek a way of avoiding the potential problem. Possibly the experimental procedure needs revision, or perhaps all that is needed is increased caution at key points. What affects precision at each step? What affects accuracy? Which steps must be done painstakingly, and which are more robust? The ideal of doing every step of every experiment as painstakingly as possible is not only unrealistic; it is a recipe for scientific progress that is so slow and inefficient that the scientist will have trouble keeping a job.
- Seek ways of streamlining the list of steps without jeopardizing the integrity and reliability of the experiment. Could some steps be done more efficiently as batch process rather than in series? Are there long wait times during which other steps can be done? Where are the natural break points for each day's work?

10) do a gedanken experiment, a thought experiment, before the actual experiment. Try to predict all of the possible outcomes of the experiment, how you would interpret them, what the weaknesses would be, and what the alternative explanations would be for each interpretation. Can the experimental design be changed to provide a more diagnostic, less ambiguous interpretation of each result? Pretend that you are a reviewer, intent on finding a fatal flaw because the results are contrary to prevailing theory. Remember that the *gedanken* experiment takes up a small fraction of the time of the actual experiment. When writing up results, we usually wish that we had done some part of the experiment differently; often a more careful *gedanken* experiment would have helped.

11) avoid last-minute changes to the experiment, unless you have carefully thought through all of their possible implications.

12) use identification labels on samples. Use indelible ink for labeling, and assure that either labels cannot be lost or labeling is redundant. If samples are in containers and if feasible, label both the sample and container.

13) take methodical, detailed notes during the experiment.

- Do not trust your memory for anything. Remember that you may come back to these notes months later, long after short-term memory of temporarily obvious factors has faded.
- Do not use scraps of paper. Ideally, use a lab notebook; some researchers say that a lab notebook is essential. At least, use dated sheets of paper and either a 3-ring binder or manila folder.
- Sketches may be useful.
- Decide whether or not to use a checklist during the experiment.
- Prepare and use a standard form for routine measurements, both to facilitate later analysis and to assure that all relevant information is recorded.

[Larson, 1989]

- Note times of steps, sample ID's, experimenter (if more than one), and anything else that remotely could be considered a variable (e.g., source and grade of chemicals) in later review of the experiment.
- Note units of all data. A frequent pitfall is to assume that the units are so obvious or familiar that you could not forget them.
- Note any changes to the experimental procedure or equipment.
- Document calibrations.
- Record raw data, not just corrected data, because you may decide later to use different corrections. Record correction equations, because you may wonder later whether or not you did all corrections properly. Raw data are better than corrected data, if the corrections are untrustworthy or of unknown accuracy. For example, in using 'temperature compensated' equipment, I have been confronted with the challenge of evaluating whether the compensation actually introduced error because of its inaccurate measurement of temperature.
- Record bad data, unreliable results, and abortive experiments, using obvious flags to avoid mistaking them for trustworthy data (e.g., draw a large X through them). Add a notation on why they failed. Possibly, later analysis will show that information can be salvaged from these discards. Certainly, one wants to minimize the chances of making the same mistake twice.

- Remember in note-taking that some 'facts' assumed during the experiment may later be rejected. Associated data therefore may be inaccurate or imprecise and need unanticipated corrections. Whether these corrections are possible probably will depend on completeness of the notes.
- Flag any unexpected observations. Immediately consider whether they may indicate a problem with experimental procedure. Later, consider whether or not they can offer a new insight (Chapter 8).
- In deciding whether or not to record something, remember how cheap paper is compared to the cost of redoing an experiment.
- Similarly, paper is too cheap to warrant tiny, crowded marginal notations that might later be overlooked or found to be unreadable.
- *Keep and regularly update a backup of your most vital notes and data*. Be prepared for the eventuality of losing notes, data listings, or data files. Take steps to postpone that eventuality: do not take prime data on airplanes as checked baggage; be cautious about carrying prime data around with you routinely; both lock and back up your personal computer; keep backups in a different room from originals.

14) protect your experimental setup, experiment, and samples **from accidental damage** by yourself or others.

- Make a sign such as "Experiment in progress, do not disturb" and display it whenever the experiment is untended. I have seen incidents where using a seemingly overcautious sign *could* have prevented heartbreaking setbacks.
- When leaving a shared computer while it is number-crunching or otherwise in use, put a sheet of paper saying "In use" over the screen or keyboard.

[Larson, 1985]

I know of incidents of janitors innocently discarding:

- data or samples stored in a container that resembled a trash can;
- delicate samples wrapped in Kleenex and left on a desk;
- boxes that were *almost* empty.

15) avoid minor changes during the experiment. They separate data obtained before and after the change with a gulf of ambiguous comparison.

16) before the experiment is completed, begin **preliminary data reduction and analysis.**

- Allow time to think about what you are observing, regardless of how busy you are just collecting data.
- Rough, first-order corrections and analysis, including back-of-envelope plots, are acceptable at this stage.
- Determine whether or not you are generating unreliable data.
- Seek clues to needed improvements (e.g., finding a major unremoved variable). While avoiding minor changes, consider the advisability of restarting the experiment with a substantial improvement.
- Beware of potential bias to subsequent results caused by expectations from the preliminary analysis.
- Do not let these preliminary analyses substitute for post-experiment, systematic data reduction and analysis.
- Some experimenters or their managers find it fruitful to write progress reports regularly during the experiment.

17) handle calculations scientifically:

- Omit meaningless digits. Usually the final result will have no more significant digits than the least-accurate variable in the calculation. Carrying one superfluous digit is preferable to omitting a meaningful digit. A propagation-of-errors analysis is even better.
- Average raw data rather than final processed data, to save steps.
- Check your calculations. If using a calculator, use a different keying sequence than for the initial calculation, to avoid making the same mistake twice. If using a computer, check results with a calculator for one or two of the samples. Computers usually make no mistake or make the same mistake for every sample, if they are correctly interpreting the input format of all of the data. However, exceptions exist (e.g., calculations that work OK for data values greater than zero but not for data less than zero).
- Ask yourself whether or not the result looks reasonable. In the old slide-rule days, quick-and-dirty estimation was essential; now, this skill is rare.

Subsequent experimental steps are less relevant to the subject of experimental design and can be left to other chapters. These include: analyzing data, interpreting the experimental results, drawing conclusions, comparing these conclusions to those of other studies, and designing a modified experiment to test the conclusions.

* * *

Pitfalls of Experimental Design

> "Faulty execution of a winning combination has lost many a [chess] game on the very brink of victory. In such cases a player sees the winning idea, plays the winning

> sacrifice and then inverts the order of his follow-up moves or misses the really clinching point of his combination." [Reinfeld, 1959]
>
> When the exorcist arrived at the house, he almost immediately started upstairs to begin the exorcism. "Wait," interrupted the attending priest, "Don't you want to learn the personalities of the demons?" "There is only one," replied the exorcist. [Blatty, 1972]

Many of the potential pitfalls to optimum experimental design are obvious from earlier parts of this chapter, particularly the section, 'Tips on Experimental Design and Execution'. Most of these pitfalls, however, are manifestations of the same demon: a **rogue, or uncontrolled, variable**.

* * *

Control of Variables

Rogue variables are a frequent scientific problem. Suspect such a problem when troubleshooting equipment or an experimental setup, if none of the initial troubleshooting techniques helps. Also suspect such a problem whenever an experiment gives surprising, unexpected results. Such problems are always a nuisance, but sometimes their solution can foster scientific insight.

"The notion of a finite number of variables is an idealization" [Wilson, 1952] that is essential to practical science. Most 'relevant' variables have only a trivial influence on the phenomenon of interest. Often, they have no direct causal relationship to this phenomenon or variable, but they do have some effect on one of the primary causal variables. Such variables are second or third-order problems that are ordinarily ignored. Usually the scientific focus is on identifying and characterizing the primary causal variables -- those that have the greatest influence on the phenomenon of interest.

In the so-called ideal experiment, the investigator holds all relevant variables constant except for a single variable. This **independent variable** is deliberately varied while measuring the resulting changes in a **dependent variable**. Simplicity gives power to such experiments, but they are based on the often dubious assumption that one knows all relevant variables. Usually, we hold as many relevant variables constant as possible and cope with the non-constant variables through randomization. Unfortunately, the variables that we can control are not necessarily the ones that are most important to control.

In Chapters 2 and 3, we considered statistical techniques for quantitatively estimating the influence of variables. Here the focus is on several methods for determining whether or not a variable is crucial. Selection of the most appropriate procedure depends on feasibility and on time and effort needed to remove or measure a variable.

Common techniques for dealing with a problem variable are:

• **stabilization**: Keeping a variable constant prevents it from influencing other variables. This approach is best for variables that are a disruptive influence (e.g., voltage or temperature variations), rather than scientifically interesting. Rarely, it is feasible to monitor the problem variable, then make measurements only when it has a certain value. The technique does not work for intermittent problems.

• **standard sample**: A control or standard sample is a way of coping simultaneously with abundant uncontrollable or even unknown variables that might otherwise bias the measurements or mask the target relationship. Repeated measurements of this standard indicate how much data variability is generated by other variables. Sometimes one must accept that the observations are relative rather than absolute, because of the biasing effects of uncontrolled variables. Often, however, a suite of standards can allow calibration of the observations. They also can enable comparison to measurements by other investigators.

• **control group**: Dividing the experiment into two groups can demonstrate the effect of a variable of interest, even when many uncontrollable or unknown variables are present. Use of a control group is standard in social sciences such as psychology, but it may also be useful wherever one must cope with several uncontrolled variables.

The two groups should be as similar as possible, except that the problem variable is missing from the 'control group' and present in the 'experiment group'. Such an experiment is called a **controlled experiment**. Note that this term does not mean that the experiment is under control (almost all experiments are controlled in that sense), but that it employs experiment and control groups.

The two-group experiment described here is the simplest controlled experiment; often it is not the most efficient experiment. Multivariate experiments using a factorial design permit one to explore the possible effects of several variables and their interactions in one experiment, rather than in sequential experiments. Design of such experiments is described in most statistics books.

An essential ingredient of most controlled experiments is randomization. Random assignment of individual samples to the two groups avoids bias and permits statistical determination of confidence levels for the effect of the variable of interest. For example, drug tests routinely use a controlled experiment with randomization and double blinds: subjects are randomly assigned to receive either the drug or a placebo, and neither the subject nor the drug dispenser knows which type is received.

> Understanding the effects of acid rain on lakes has been hampered by the complexity of lake systems and the very high variability among lakes. Thus even when a lake with rising acid levels undergoes ecologic change, it is not possible to establish causality between the two.
>
> Recent experiments in Canada have been able to detect ecologic changes caused by only minor acidification. They demonstrated, for example, that acidification causes a decrease in species diversity without changing total biomass -- an observation consistent with the more general ecologic pattern that environmental extremes affect diversity but not necessarily numbers. The experiments used the following drastic technique: choose a remote region of Canada where lakes are plentiful, select a pair of environmentally similar lakes, make one of each pair a control and deliberately acidify the second lake, then monitor the changes in both [Luoma, 1992].

• **randomization**: If an experimental design randomly selects samples or randomly matches treatment to samples, then potential biasing effects of uncontrolled variables are converted into random unbiased error that can be averaged out. For example, time often is an influential variable, because instruments may drift or subtle changes may creep into the experimental setup. By randomizing the sequence of sample measurements, the investigator can prevent undetected temporal changes from biasing the result. Randomization is the most powerful tool for dealing with uncontrolled variables;

it succeeds whether or not you are aware of their presence. Randomization is less efficient that the other methods, however, because it converts bias into random noise, rather than quantifying or removing bias.

• **correlation**: If you cannot control a problem variable but can measure it, measure and record its value at each data measurement. Later, crossplot the variable of interest versus this problem variable. This technique succeeds even if the relationship between variables is nonlinear. It has disadvantages (Chapter 3): both types of measurement may change as a function of time, leading to a noncausal correlation, or a time lag may obscure the relationship.

• **artificial variation**: Deliberately change the problem variable by more than it is likely to change normally, in order to estimate the conditions under which this variable is prominent, as well as its maximum possible effect. The advantage of this technique is its ability to detect effects that are ordinarily subtle, by exaggerating them. The main disadvantage is that ordinarily trivial effects can be misinterpreted as disruptive. When the relationship between two variables is highly nonlinear, artificial variation is a poor predictor of the normal relationship.

> When Irving Langmuir was trying to develop a new light bulb, he knew that ideally its interior should have a perfect vacuum. Faced with the impossibility of attaining that ideal, Langmuir deliberately added different gases to assess their effects. He discovered the gas-filled (fluorescent) light. Langmuir [1928] said,
>
> "This principle of research I have found extremely useful on many occasions. When it is suspected that some useful result is to be obtained by avoiding certain undesired factors, but it is found that these factors are very difficult to avoid, then it is a good plan to increase deliberately each of these factors in turn so as to exaggerate their bad effects, and thus become so familiar with them that one can determine whether it is really worthwhile avoiding them."
>
> Another example: if you suspect that changes in equipment readings are caused by a temperature-sensitive electronic component, remove the equipment housing and blast various components with either a heat gun (e.g., a hair dryer) or coolant gas, while monitoring equipment readings.

An alternative to artificial variation is to investigate naturally occurring extreme points. The advantage is the same: maximizing an ordinarily subtle effect, to evaluate its potential impact.

> Numerous studies of type-Ia supernovae during the past several years have shown a consistent pattern of increasing redshift with decreasing apparent magnitude (i.e., greater speed at greater distance) that implies that the expansion of the universe is accelerating. This unexpected conclusion was not compelling, however. The observed pattern could also be produced by dust or chemical evolution. A single new data point, from a supernova with a redshift of 1.7, far beyond the 0.3-0.9 range of previous data, excludes the alternative ideas and confirms that the universe is accelerating [Schwarzschild, 2001].

• **sequential removal**: When more than one variable may be influential, remove the dominant one and look at the effect of the next variable on the data of interest. Then remove this variable as well, so that possible effects of additional variables can be examined. This technique works only when the problem variables are controllable and their relative importance can be estimated. Nevertheless, it

can be quite valuable or even essential. For example, if you think that variables X_1, X_2, and X_3 may be disrupting your readings of D as a function of A, then temporarily keep A constant and record variations of D, X_1, X_2, and X_3. At this reconnaissance stage, these problem variables need not be controllable. If they are controllable, however, factorial design is a more powerful experimental technique: it allows us to both isolate and quantify the influence of these variables. A related approach is the method of residuals (Chapter 3): measure variations caused by the dominant variable, remove its estimated effects, then compare data residuals to second-order variables.

> Studies of the causes of spread of the AIDS disease long ago established that most U.S. cases are attributable to homosexual or intravenous transmission. But does heterosexual transmission occur, and if it does, how abundant is it? One technique to examine these questions is clearly biased, yet it is apparently the best available. *Any* AIDS instance that could be either homosexually or intravenously transmitted is attributed to those origins rather than to heterosexual transmission, regardless of the relative abundance of heterosexual versus other encounters. Only cases in which homosexual or intravenous transmission are impossible are attributed to heterosexual transmission. Because (we think) heterosexual transmission is much less likely per encounter than are other forms of transmission, this accounting bias toward the dominant variables is considered to be acceptable [Hilts, 1992].

* * *

Problem: the Noisy Widgetometer

You need to measure some widgets on your new high-precision widgetometer. Before starting, however, you prudently run some standard samples and find that the precision and accuracy are far below what is advertised. In desperation, you connect the widgetometer to a chart recorder and let it run for 24 hours, obtaining the record in Figure 21. How do you interpret this record, and what techniques and experimental designs could you use to deal with the problem?

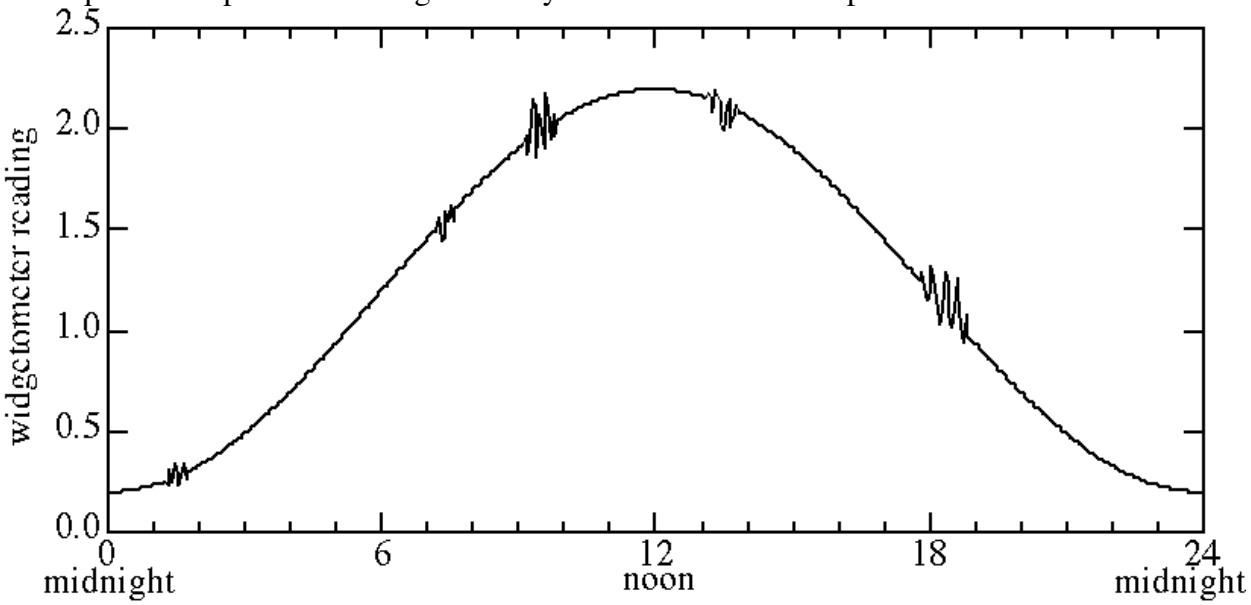


Figure 21. Chart recording of widgetometer readings over a 24-hour period. Note the long-term instrument drift and occasional noise spikes.

Answer: The instrument is exhibiting a daily drift plus occasional noise spikes. First priority is to try to identify and remove the drift. Second priority is to minimize the disruptive effects of any residual drift. Possible sources of daily cycles are daily temperature fluctuations and voltage fluctuations; try changing these and alternative variables substantially while running the chart recorder. If you can identify the variable, try to prevent it from affecting your measurements (e.g., voltage regulator), or quantify the relationship and monitor that variable during all measurements, so that you can apply a correction.

If the cause of the daily variations is unknown or unmeasurable, choose an experimental design that minimizes its effect. The most obvious is to take either a zero reading or a calibration-standard measurement along with each widget measurement, depending on whether drift is in zeroing or in sensitivity, respectively.

The cause of the intermittent noise spikes is likely to be quite elusive. Because they are sudden and short-lived, they could make some measurements much less accurate than most, and they could affect only one of a paired measurement. One approach would be to measure the zero or standard both immediately before and immediately after the sample. If the two zero/standard measurements differ by more than a predetermined threshold, reject this measurement set and do another.

* * *

Computation and Information Handling

Computers are wonderful productivity enhancers. Whether for word processing, figure preparing, calculating, or extracting the most information from data, computers are essential to modern science. When I was a young scientist, I would give a draft manuscript to a secretary for typing, have one or at most two rounds of revisions, and submit it. I would give a roughed-out figure to a draftsperson, flag the most glaring drafting errors for revision, and submit it. Now I do my own typing and drafting, and I do dozens of revisions! The process as a whole may be slower, but the final product is certainly more polished.

Basic computer literacy for scientists includes proficiency in all of the following:

- an operating system (Windows®, Macintosh®, Unix®, or Linux®);
- word processing (e.g., Word® or Word Perfect®);
- spreadsheet analysis (e.g., Excel®); and
- a web browser (Netscape® or Internet Explorer®).

Most scientists also need one or more of the following:

- a graphics program (e.g., Kaleidagraph®);
- presentation software for slides and transparencies (e.g., PowerPoint®);
- image handling software (Photoshop® or Canvas®); and
- a statistical package (WinStat®, MINITAB®, SAS®, or SYSTAT®).

* * *

For some kinds of computation, speed is power. The current generation of computers is capable of solving more complex problems, involving more dimensions or variables, than were feasible even five years ago. The fastest vector machines, such as the Cray, are approaching their ultimate speed limits. Parallel processing, in contrast, is not bound by those limitations. Today's largest computational tasks are massive because of the size of matrices or datasets, rather than because of the number of different kinds of computations. Such problems are well suited to parallel processing. The CM-2 Connection Machine, introduced in 1987, is an example of massive parallel processing: ef-

fectively it is 65,536 processors, each capable of associating both with its neighbors and with an individual data point [Boghosian, 1990].

Giant and expensive parallel computers are an exception to the overall trend toward small personal computers. An emerging alternative to the parallel-processor supercomputers is distributed computing. Ten to twenty high-performance workstations (e.g., Suns) are used simultaneously, via message-passing software and a local area net, to run different parts of the same problem. Effectively, the workstations become a virtual parallel computer, and they do so at night or in the background so that their metamorphosis disturbs nobody.

The acceleration of processing capability is generating opportunities for scientific projects that were previously impossible. Modeling can encompass complex systems (e.g., econometric models) and three dimensions (e.g., global climate models). Inversion can involve huge datasets (e.g., Human Genome Project) and three-dimensional, non-invasive tomographic imaging (e.g., CT scans, tomography of Earth's interior). Image analysis of immense datasets is feasible (e.g., astronomy).

For most scientists, personal computers are sufficient and in fact superior to supercomputers. Scientists value control, and having one's own computer, with a simple enough operating system to eliminate system managers, provides that control. Indeed, the major obstacle to further expansion of distributed computing may be the reluctance of individuals to relinquish a fraction of their supervision of their own computers.

* * *

Neither large nor small computers have removed the need for a vintage type of scientific calculation: back-of-the-envelope calculations. Computers have 8-digit or more accuracy, but the back-of-the-envelope calculation recognizes that the reliability of many calculations depends instead on huge uncertainty in one or two of the needed variables. Even the most advanced computer is naïve about pivotal concerns such as estimation and the difference between random and systematic errors. The scientist must provide the missing sophistication, either explicitly in a back-of-the-envelope calculation or implicitly in the data input to a computer algorithm. Chapter 2 addresses some of these concerns.

> Late at night, sharing a Coke, feeling guilty about its 130 calories, my wife and I recalled the cryogenic diet, which we had seen long ago in a Journal of Irreproducible Results. Total dietary impact is not 130 calories, but 130 calories *minus* the calories required to heat the liquid from ice-cold (0°C) to body temperature (~35°C). A calorie, I knew from recently preparing an Oceanography lecture, is the heat required to raise 1 cc of water 1°C. A back-of-an-envelope calculation showed the benefit of a 12-ounce ice-water diet:
>
> 12 oz x ~35g/oz x 1 cc x 35°C x 1 calorie/cc°C ≈ 13,000 calories!
>
> We realized that a 'Popsicle diet' (2 6-oz Popsicles) would be even better: 13,000 calories for warming from 0°C to 35°C, plus 32,000 calories (400 cc x 80 calories/cc) heat of transformation from ice to water! Clearly, there was a problem, and not one that a calculator or computer could solve. Days later, my wife found the answer: oceanographers use 'small' calories (1 g heated 1°C), but dietary calories are 'large' calories (1 kg heated 1°C). Neither anticipated the loss of sleep that a factor of 1000 could cause in a couple of hapless scientists.

When using calculators and personal computers, extra attention is needed concerning significant digits. Significant digits, or significant figures, are an implicit statement about the precision of a measurement. In general, a measurement of completely unknown precision is virtually worthless.

Ideally, each measurement given in a publication would be accompanied by a calculated estimate of its precision. Precision estimates, however, generally require replicate measurements, which may not be available. The use of significant digits may have to suffice. *The number of significant digits is equal to the number of digits that are reliably known, ignoring leading zeros*.

Although the rules concerning significant digits are simple, few of the current software packages honor them. Some follow the conservative approach of assuming that all digits are significant (e.g., 1÷3=0.333333...). Some strip off trailing zeros whether or not they are significant; for example, a series of numbers accurate to ±0.01 might appear as 1.14, 1.1, 1.07, and 1. Most maintain a user-selectable constant number of digits to the right of the decimal place. None of these conventions is appropriate for publication.

* * *

The word *computer* is no longer appropriate. The proportion of computer usage devoted to computation is steadily decreasing. Many of the recent computer developments have had little to do with computation. Of particular interest to scientists is the extent to which computer networking is revolutionizing information handling.

Efficient information handling has always been an essential aspect of scientific method. Even the early days of science had more observations -- mostly irrelevant -- than a mind could encompass; an example is Leonardo da Vinci's quicksilver mind and notes. Today, information handling is a mantra of our technological society. Are we witnessing another transient enthusiasm, or are we truly challenged to adapt or be left behind?

Research faces two information problems -- locating and organizing relevant information. These problems are relatively minor in the course of one's own experiments, although they certainly are felt while writing up results. The real hurdle is in dealing with the vast published literature. All memories are fallible, especially my own. Where was I?

The first step in information handling is skimming or digesting a scientific publication. These days, we usually have a personal copy of the paper rather than the library's, so we are free to mark up the paper with underlines and marginal comments. To organize information from several papers, many people simply group and rescan a stack of reprints. Others prefer to take notes, either on a pad or laptop. A virtue of the latter is easy reorganization, because association is essential to pattern recognition. Furthermore, typing is faster than writing, and the *Find* command is a great time saver.

Ambitious schemes for information handling tend to fail. First the scientist falls behind in entering data into the system. Later, the backlog is so great that the system atrophies.

Is information handling by computers more efficient than by scientists? For straightforward sorting, bookkeeping, and information archiving, the answer is yes. The quantity, or content, of science is doubling every five years, so the need for efficient data handling is undoubted. Use of the Internet and the World Wide Web is growing exponentially, and every scientist faces the question of how much to employ these valuable tools. Both publications and published data are becoming more available on the Internet. We can anticipate, as a result, increased awareness of relevant publications and more analyses of data by individuals other than the one who collected the data. Where better to exploit the Information Age than in the quest for answers to scientific questions?

Whenever one develops a hypothesis, the first step is to see whether or not it survives the test of existing data. If we decide that none of the previous relevant experiments was appropriately designed to provide a diagnostic test of the hypothesis, only then do we conduct a new experiment. Scientific progress does not imply, however, that the same person who generates hypotheses tests them. Already, many scientists are tapping the information river to produce papers that present no

new data. Instead, they use a variety of published data to test hypotheses and develop syntheses. For the experienced Internet traveler, an easy path to scientific productivity is to read a newly proposed hypothesis and then extract data from the Internet to test it.

Scientists who rarely employ the Web may find that they are left behind, even in exploring their own hypotheses. Other scientists, however, are falling victim to the other extreme -- net surfing. Too often, Internet and the Web are used just for browsing rather than for goal-oriented information retrieval. The hallway refrain is "Did you see . . .?" And there is much to browse. Some scientists respond by devoting weeks to developing their own web pages. I, who published this book on-line rather than on paper, am in a poor position to criticize. Perhaps there is no cause for concern. Presumably, those whom I have watched wandering off into the Web will return.

Chapter 6: The Myth of Objectivity

Scientists seek concepts and principles, not subjective perspectives. Thus, we cling to a myth of objectivity: that direct, objective knowledge of the world is obtainable, that our preconceived notions or expectations do not bias this knowledge, and that this knowledge is based on objective weighing of all relevant data on the balance of critical scientific evaluation. In referring to objectivity as a myth, I am not implying that objectivity is a fallacy or an illusion. Rather, like all myths, objectivity is an ideal -- an intrinsically worthwhile quest.

> "One aim of the physical sciences has been to give an exact picture of the material world. One achievement of physics in the twentieth century has been to prove that that aim is unattainable.
>
> "There is no absolute knowledge. . . All information is imperfect. We have to treat it with humility." [Bronowski, 1973]

In this chapter we first will examine several case studies that demonstrate ways in which perception is much less objective than most people believe. Our primary means of scientific perception is visual: 70% of our sense receptors are in the eyes. Thus our considerations of perception will focus particularly on visual perception. We then will examine theories of how perception operates, theories that further undermine the fantasy of objectivity. These perspectives allow us to recognize the many potential pitfalls of subjectivity and bias, and how we can avoid them. Finally, we will address a critical question: can a group of subjective scientists achieve objective scientific knowledge?

* * *

Perception: Case Studies

> "Things are, for each person, the way he perceives them." [Plato, ~427-347 B.C., b]

What do the following topics have in common: football games, a car crash, flash cards, a capital-punishment quiz, relativity, and quantum mechanics? The study of each provides insight into the perception process, and each insight weakens the foundation of objectivity.

I was never much of a football fan. I agree with George Will, who said that football combines the two worst aspects of American life: it is violence punctuated by committee meetings. Yet, I will always remember two games that were played more than 20 years ago. For me, these games illuminate flaws in the concept of objective perception, suggesting instead that: 1) personal perception can control events, and 2) perceptions are, in turn, controlled by expectations.

It was a high school game. The clock was running out, and our team was slightly behind. We had driven close to the other team's goal, then our quarterback threw a pass that could have given us victory; instead the pass was intercepted. Suddenly the interceptor was running for a touchdown, and our players merely stood and watched. All of our players were at least 20 yards behind the interceptor.

Though it was obvious to all of us that the attempt was hopeless, our star halfback decided to go after him. A few seconds later he was only two yards behind, but time had run out for making up the distance -- the goal was only 10 yards ahead. Our halfback dived, and he chose just the right moment. He barely tapped his target's foot at the maximum point of its backward lift. The halfback

cratered, and the interceptor went on running. But his stride was disrupted, and within two paces he fell -- about two yards short of a touchdown.

I prefer to think that our team was inspired by this event, that we stopped the opponents' advance just short of a touchdown, and that we recovered the ball and drove for the winning touchdown. Indeed, I do vaguely remember that happening, but I am not certain. I remember that the halfback went on to become an All Star. Of this I am certain: I could read the man's thoughts, those thoughts were "No, damn it, I refuse to accept that," and willpower and a light tap at exactly the right moment made an unforgettable difference.

* * *

The game that affected me most I never saw. I read about it in a social anthropology journal article 27 years ago. The paper, called 'They saw a game,' concerned a game between Harvard and perhaps Yale or Dartmouth. The authors, whose names I don't remember, interviewed fans of both teams as they left the game.

Everyone agreed that the game was exceedingly dirty, and the record of the referees' called fouls proves that assertion. Beyond that consensus, however, it was clear that fans of the two teams saw two different games. Each group of fans saw the rival team make an incredibly large number of fouls, many of which the referees 'missed'. They saw their own team commit very few fouls, and yet the referees falsely accused their team of many other fouls. Each group was outraged at the bias of the referees and at the behavior of the other team.

The authors' conclusion was inescapable and, to a budding scientist imbued with the myth of scientific objectivity, devastating: *expectations exert a profound control on perceptions*. Not invariably, but more frequently than we admit, we see what we expect to see, and we remember what we want to remember.

Twenty-three years later, I found and reread the paper to determine how accurate this personally powerful 'memory' was. I have refrained from editing the memory above. Here, then, are the 'actual' data and conclusions, or at least my current interpretation of them.

In a paper called 'They saw a game: a case study,' Hastorf and Cantril [1954] analyzed perceptions of a game between Dartmouth and Princeton. It was a rough game, with many penalties, and it aroused a furor of editorials in the campus newspapers and elsewhere, particularly because the Princeton star, in this, his last game for Princeton, had been injured and was unable to complete the game. One week after the game, Hastorf and Cantril had Dartmouth and Princeton psychology students fill out a questionnaire, and the authors analyzed the answers of those who had seen either the game or a movie of the game. They had two other groups view a film of the game and tabulate the number of infractions seen.

The Dartmouth and Princeton students gave discrepant responses. Almost no one said that Princeton started the rough play; 36% of the Dartmouth students and 86% of the Princeton students said that Dartmouth started it; and 53% of the Dartmouth students and 11% of the Princeton students said that both started it. But most significantly, out of the group who watched the film, the Princeton students saw twice as many Dartmouth infractions as the Dartmouth students did.

Hastorf and Cantril interpreted these results as indicating that, when encountering a mix of occurrences as complex as a football game, we experience primarily those events that fulfill a familiar pattern and have personal significance.

> Hastorf and Cantril [1954] conclude: "In brief, the data here indicate that there is no such 'thing' as a 'game' existing 'out there' in its own right which people merely 'observe.'"

Was my memory of this paper objective, reliable, and accurate? Apparently, various aspects had little lasting significance to me: the teams, article authors, the question of whether the teams were evenly guilty or Dartmouth was more guilty of infractions, the role of the Princeton star in the debate, and the descriptive jargon of the authors. What was significant to me was the convincing evidence that the two teams 'saw' two different games and that these experiences were related to the observers' different expectations: I remembered this key conclusion correctly.

I forgot the important fact that the questionnaires were administered a week after the game rather than immediately after, with no attempt to distinguish the effect of personal observation from that of biasing editorials. As every lawyer knows, immediate witness accounts are less biased than accounts after recollection and prompting. I forgot that there were also two groups who watched for infractions as they saw a film, and that these two groups undoubtedly had preconceptions concerning the infractions before they saw the film.

The experiment is less convincing now than it was to me as an undergraduate student. Indeed, it is poorly controlled by modern standards, yet I think that the conclusions stand unchanged. The pattern of my selective memory after 23 years is consistent with these conclusions.

I have encountered many other examples of the subjectivity and bias of perception. But it is often the first unavoidable anomaly that transforms one's viewpoints. For me, this football game -- although hearsay evidence -- triggered the avalanche of change.

Hastorf and Cantril [1954] interpreted their experiment as evidence that "out of all the occurrences going on in the environment, a person selects those that have some significance for him from his own egocentric position in the total matrix." Compare this 'objective statistical experimental result' to the much more subjective experiment and observation of Loren Eiseley [1978]:

> "Curious, I took a pencil from my pocket and touched a strand of the [spider] web. Immediately there was a response. The web, plucked by its menacing occupant, began to vibrate until it was a blur. Anything that had brushed claw or wing against that amazing snare would be thoroughly entrapped. As the vibrations slowed, I could see the owner fingering her guidelines for signs of struggle. A pencil point was an intrusion into this universe for which no precedent existed. Spider was circumscribed by spider ideas; its universe was spider universe. All outside was irrational, extraneous, at best raw material for spider. As I proceeded on my way along the gully, like a vast impossible shadow, I realized that in the world of spider I did not exist."

Stereotypes, of football teams or any group, are an essential way of organizing information. An individual can establish a stereotype through research or personal observation, but most stereotypes are unspoken cultural assumptions [Gould, 1981]. Once we accept a stereotype, we reinforce it through what we look for and what we notice.

The danger is that a stereotype too easily becomes prejudice -- a stereotype that is so firmly established that we experience the generalization rather than the individual, regardless of whether or not the individual fits the stereotype. When faced with an example that is inconsistent with the stereotype, the bigot usually dismisses the example as somehow non-representative. The alternative is acceptance that the prejudice is imperfect in its predictive ability, a conclusion that undermines one's established world-view [Goleman, 1992c].

Too often, this process is not a game. When the jury verdict in the O.J. Simpson trial was announced, a photograph of some college students captured shock on every white face, joy on every

black face; different evidence had been emphasized. The stakes of prejudice can be high, as in the following example from the New York Times.

> "JERUSALEM, Jan. 4 -- A bus driven by an Arab collided with a car and killed an Israeli woman today, and the bus driver was then shot dead by an Israeli near the Gaza Strip.
>
> "Palestinians and Israelis gave entirely different versions of the episode, agreeing only that the bus driver, Mohammed Samir al-Katamani, a 30-year-old Palestinian from Gaza, was returning from Ashkelon in a bus without passengers at about 7 A.M. after taking families of Palestinian prisoners to visit them in jail.
>
> "The bus company spokesman, Mohammed Abu Ramadan, said the driver had accidentally hit the car in which the woman died. He said the driver became frightened after Israelis surrounded the bus and that he had grabbed a metal bar to defend himself.
>
> "But Moshe Caspi, the police commander of the Lachish region, where the events took place, said the driver had deliberately rammed his bus into several cars and had been shot to death by a driver of one of those vehicles.
>
> "The Israeli radio account of the incident said the driver left the bus shouting 'God is great!' in Arabic and holding a metal bar in his hand as he tried to attack other cars." [Ibrahim, 1991]

* * *

> "What a man sees depends both upon what he looks at and also upon what his previous visual-conceptual experience has taught him to see." [Kuhn, 1970]

The emotional relationship between expectation and perception was investigated in an elegantly simple and enlightening experiment by Bruner and Postman [1949]. They flashed images of playing cards in front of a subject, and the subject was asked to identify the cards. Each card was flashed several times at progressively longer exposures. Some cards were normal, but some were bizarre (e.g., a red two of spades).

Subjects routinely identified each card after a brief exposure, but *they failed to notice the anomaly*. For example, a red two of spades might be identified as a two of spades or a two of hearts. As subjects were exposed more blatantly to the anomaly in longer exposures, they began to realize that something was wrong but they still had trouble pinpointing the problem. With progressively longer exposures, the anomalous cards eventually were identified correctly by most subjects. Yet nearly always this period between recognition of anomaly and identification of anomaly was accompanied by confusion, hesitation, and distress. Kuhn [1970] cited a personal communication from author Postman that even he was uncomfortable looking at the bizarre cards. Some subjects never were able to identify what was wrong with the cards.

The confusion, distress, and near panic of attempting to deal with observations inconsistent with expectations was eloquently expressed by one subject:

> "I can't make the suit out, whatever it is. It didn't even look like a card that time. I don't know what color it is now or whether it's a spade or a heart. I'm not even sure now what a spade looks like. My God!"

* * *

Let us now consider a critical aspect of the relationship between expectation and perception: how that relationship is reinforced. Lord et al. [1979] investigated the evolution of belief in an hypothesis. They first asked their subjects to rate how strongly they felt about capital punishment. Then they gave each subject two essays to read: one essay argued in favor of capital punishment

and one argued against it. Subsequent quizzing of the subjects revealed that they were less critical of the essay consistent with their views than with the opposing essay. This result is an unsurprising confirmation of the results of 'They saw a game' above.

The surprising aspect of Lord et al.'s [1979] finding was this: reading the two essays tended to reinforce a subject's initial opinion. Lord et al. [1979] concluded that examining mixed, pro-and-con evidence further polarizes initial beliefs. This conclusion is particularly disturbing to scientists, because we frequently depend on continuing evaluation of partially conflicting evidence.

In analyzing this result, Kuhn et al. [1988] ask the key question: what caused the polarization to increase? Was it the consideration of conflicting viewpoints as hypothesized by Lord et al. [1979] or was it instead the incentive to reconsider their beliefs? Kuhn et al. [1988] suspect the latter, and suggest that similar polarization might have been obtained by asking the subjects to write an essay on capital punishment, rather than showing them conflicting evidence and opinions. I suspect that both interpretations are right, and both experiments would increase the polarization of opinions. Whether one is reading 'objective' pro-and-con arguments or is remembering evidence, one perceives a preponderance of confirming evidence.

Perception strengthens opinions, and perception is biased in favor of expectations.

* * *

Though the preceding case studies demonstrate that perception is much less objective and much more belief-based than we thought, they allow us to maintain faith in such basic perceptual assumptions as time and causality. Yet the next studies challenge even those assumptions.

> "Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality."

With these stunning initial words, the Russo-German mathematician Hermann Minkowski [1908] began a lecture explaining his concept of *space-time*, an implication of Albert Einstein's 1905 concept of special relativity.

Einstein assumed two principles: relativity, which states that no conceivable experiment would be able to detect absolute rest or uniform motion; and that light travels through empty space with a speed *c* that is the same for all observers, independent of the motion of its source. Faced with two incompatible premises such as universal relative motion yet absolute motion for light, most scientists would abandon one. In contrast, Einstein said that the two principles are "only apparently irreconcilable," and he instead challenged a basic assumption of all scientists -- that time is universal. He concluded that the simultaneity of separated events is relative. In other words, if two events are simultaneous to one observer, then they are not simultaneous to a second observer at a different location. Clocks in vehicles going at different speeds do not run at the same speed.

Although Einstein later relaxed the assumption of constant light velocity when he subsumed special relativity into general relativity in 1917, our concept of an objective observer's independence from what is observed was even more shaken. Space and time are, as Minkowski suggested, so interrelated that it is most appropriate to think of a single, four-dimensional, space-time. Indeed, modern physics finds that some atomic processes are more elegant and mathematically simple if we assume that time can flow either forward or backward. Gravity curves space-time, and *observation depends on the motion of the observer*.

* * *

Even if, as Einstein showed, observation depends on the motion of the observer, cannot we achieve objective certainty simply by specifying both? Werner Heisenberg [1927], in examining the

implications of quantum mechanics, developed the principle of indeterminacy, more commonly known as "the Heisenberg uncertainty principle." He showed that indeterminacy is unavoidable, because the process of observation invariably changes the observed object, at least minutely.

The following thought experiments demonstrate the uncertainty principle. We know that the only way to observe any object is by bouncing light off of it. In everyday life we ignore an implication of this simple truth: bouncing light off of the object must impart energy and momentum to the object. Analogously, if we throw a rock at the earth we can ignore the fact that the earth's orbit is minutely deflected by the impact. But what if we bounce light, consisting of photons, off of an electron? Photons will jolt the electron substantially. *Our observation invariably and unavoidably affects the object being observed*. Therefore, we cannot simultaneously know both where the electron is and what its motion is. Thus we cannot know exactly where the particle will be at any time in the future. The more accurately we measure its location, the less accurately can we measure its momentum, and vice versa.

> "Natural science does not simply describe and explain nature; . . . it describes nature as exposed to our method of questioning." [Heisenberg, 1958]

Our concepts of reality have already been revised by Einstein, Bohr, Heisenberg, and other atomic physicists; further revisions seem likely. We now know that observations are relative to the observer, that space is curved and time is relative, that observation unavoidably affects the object observed, that probability has replaced strict determinism, and that scientific certainty is an illusion. Is the observational foundation of science unavoidably unreliable, as some non-scientists have concluded?

Seldom can we blame uncontrolled observer-object interaction on atomic physics. The problem may be unintentional, it may be frequent (see the later section on 'Pitfalls of Subjectivity'), but it is probably avoidable. In our quest for first-order phenomena (such as controls on the earth's orbit), we can safely neglect trivial influences (such as a tossed rock).

Scientists are, above all, pragmatists. In practice, those of us who are not theoretical particle physicists or astronomers safely assume that time is absolute, that observation can be independent of the object observed, and that determinism is possible. For the vast majority of experimental situations encountered by scientists, these assumptions, though invalid, are amazingly effective working hypotheses. If we are in error by only one quantum, it is cause for celebration rather than worry. The fundamental criterion of science is, after all, what works.

[Larson, 1980]

We cannot, however, cling to the comfortable myth of detached observation and impartial evaluation of objectively obtained evidence. The actual process is much more complex and more human (Figure 22):

> Expectations, rooted in previous experience or in stereotypes, exert a hidden influence – usually even a control – on both our perceptions and our evaluation of evi-

dence. We tend to overlook or discount information that is unexpected or inconsistent with our beliefs. Anomaly, once recognized, can transform our perspectives fundamentally, but not without an emotional toll: "My God!" was the subject's response, when the stakes were only recognition of a playing card!

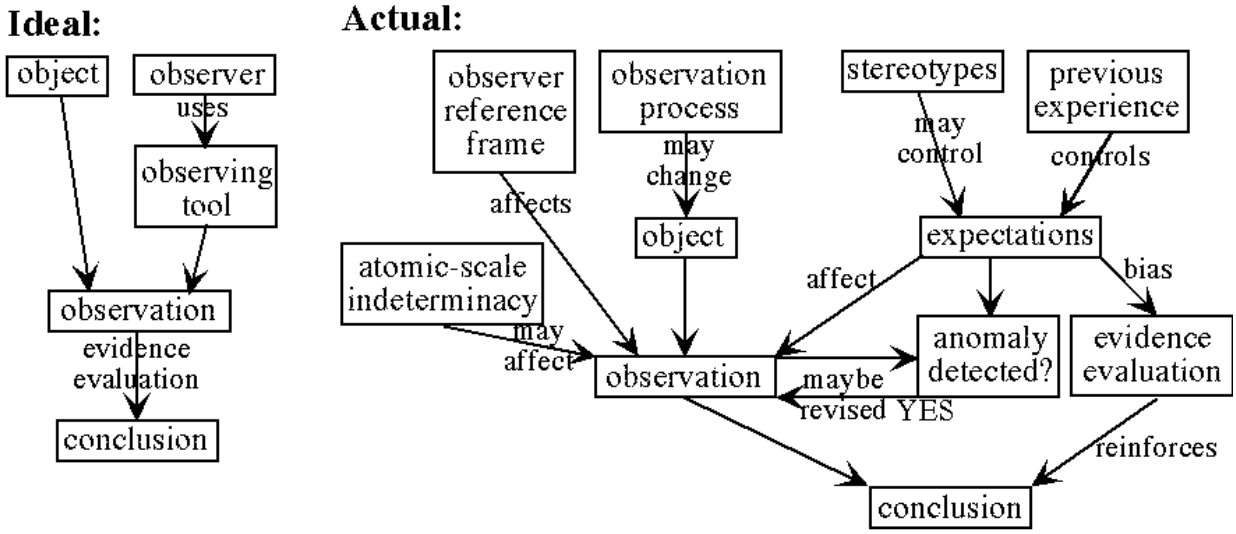


Figure 22. Flowcharts of ideal and actual interactions concerning experimental observations.

"Twenty men crossing a bridge,
Into a village,
Are twenty men crossing twenty bridges,
Into twenty villages."
[Wallace Stevens, 1931]

The scientific challenge is to use admittedly subjective means, coupled with imperfect assumptions, yet achieve an 'objective' recognition of patterns and principles. To do so, we must understand the limitations of our methods. We must understand how the perception process and memory affect our observations, in order to recognize our own biases.

* * *

Perception, Memory, and Schemata

Perception and memory are not merely biased; even today, they are only partially understood. In the mid-17th century René Descartes took the eye of an ox, scraped its back to make it transparent, and looked through it. The world was inverted. This observation confirmed Johannes Kepler's speculation that the eye resembles a camera, focusing an image on its back surface with a lens [Neisser, 1968]. Of course, 'camera' had a different meaning in the 17th century than it does today; it was a black box with a pinhole aperture, and it used neither lens nor film. Nevertheless, the analogy between camera and eye was born, and it persists today.

If the eye is like a camera, then is memory like a photograph? Unfortunately it is not; the mistaken analogy between memory and photographs has delayed our understanding of memory. The eye does not stand still to 'expose' an image; it jumps several times per second, jerkily focusing on different regions. The fovea, the portion of eye's inner surface capable of the highest resolution,

sharply discerns only a small portion of the field of view. A series of eye jerks constructs a composite, high-resolution image of the interesting portion of the visual field.

Consciousness and memory do not record or notice each jerky 'exposure'. Indeed, if we were to make a motion picture that jumped around the way the eye does, the product would be nerve-wracking or nauseating. Our vision does not bother us as this movie would, because attention controls the focus target; attention does not passively follow the jumps.

A closer analogy to functioning of the eye and mind is the photomosaic, a composite image formed by superimposing the best parts of many overlapping images. To make a photomosaic of a region from satellite photographs, analysts do not simply paste overlapping images. They pick the best photo for each region, discarding the photos taken during night or through clouds. Perhaps the eye/mind pair acts somewhat similarly; from an airplane it can put together an image of the landscape, even through scattered small clouds. Furthermore, it can see even with stroboscopic light. Both motion pictures and fluorescent light are stroboscopic; we do not even notice because of the high frequency of flashes.

Julian Hochberg suggested a simple exercise that offers insight into the relationship of vision to memory: remember how many windows are on the front of your house or apartment house [Neisser, 1968]. You may be able to solve this problem 'analytically', by listing the rooms visible from the front and remembering how many windows face front in each room. Instead, you probably will use more obvious visualization, creating a mental image of the front of your house, and then scanning this image while counting windows. This mental image does not correspond to any single picture that the eye has ever seen. There may be no single location outside your house where you could stand and count every window; trees or bushes probably hide some windows.

During the last thousand years, various cultures have grappled with the discrepancy between construct and reality, or that between perspective and reality. More than six centuries before Descartes' experiment with the eye of an ox, Arab scientists were making discoveries in optics. Alhazen, who wrote *Optics* [~1000 A.D.], realized that vision consists of light reflecting off of objects and forming a cone of light entering the eye. This first step toward an understanding of perspective was ignored by western scientists and artists. Artists tried to paint scenes as they are rather than as they appear to the eye. Beginning in the 15th century and continuing into the 16th century, artists such as Filippo Brunelleschi began to deliberately use Arab and Greek theories of perspective to make paintings appear more lifelike.

The perspective painting is lifelike in its mimicry of the way an eye or camera sees. In contrast, the older attempts to paint objects 'as they are' are more analogous to mental schemata. In this sense, the evolution of physics has paralleled that of art. Newton described dynamics as a picture of 'how things are', rather than how they appear to the observer. In contrast, Einstein demonstrated that we can only know things from some observer's perspective.

* * *

The electrical activity in the brain is not designed to store a photograph in a certain location, the way that one can electrically (or magnetically) store a scanned image in a computer. True, visual imaging is generally located in one part of the brain called the visual cortex, but removal of any small part of the brain does not remove individual memories and leave all others unchanged; it may weaken certain types of memories. Memories are much more complex than mental images. They may include visual, auditory, and other sensual data along with emotional information. Recollection involves their simultaneous retrieval from different parts of the brain. A memory consists of mental electrical activity, more analogous to countless conductors in parallel than to a scanned image. However, this analogy offers little or no real insight into how the mind/brain works.

A more illuminating perspective on both perception and memory is the concept of **schemata**, the meaningful elements of our visual environment. You or I can identify a friend's face in a group photo in seconds or less. In contrast, if you had to help someone else find your friend by describing the friend's characteristics, identification would be dilatory and uncertain. Evolutionary pressure favors fast identification: survival may depend on rapid recognition of predators. Obviously our pattern recognition techniques, whether of predators or friends, are extremely efficient and unconscious. Schemata achieve that needed speed.

Schemata are the building blocks of memories and of pattern recognition. An individual schema is somewhat like one of Plato's forms: an idealized representation of an object's essence. It may not match any specific object that we have seen. Instead, it is a composite constructed from our history of observing and recognizing examples of the object. A group of neural pathways fires during our first exposure to the object, thereby becoming associated with the object. Later recognition of the object triggers re-firing along these paths.

Schemata are somewhat like the old rhyme,

> "Big fleas have little fleas
> On their backs to bite 'em,
> And little fleas have smaller fleas
> And so *ad infinitum*."

Even a schema as 'simple' as *pencil* is made up of many other schemata: textural (e.g., hard), morphological (e.g., elongated, cylindrical, conical-ended), and especially functional. And an individual schema such as *cylindrical* is composed of lower-level schemata. Another example of a schema is a musical composition, composed of lower-level patterns of a few bars each, and -- at the lowest level -- notes.

It all sounds hopelessly cumbersome, but the process is an elegantly efficient flow of electrical currents in parallel. Identification of a pattern does not require identification of all its elements, and identification of a higher-level schema such as *pencil* does not await identification of all the lower-level schemata components. If I see a snake, then I do not have the luxury of a casual and completely reliable identification: the schemata *sinuous*, *cylindrical*, *6" to 6'* (however that information is stored as schema), and *moving* may trigger a jump response before I realize why I am jumping. I've seen a cat jump straight up into the air on encountering a sinuous stick.

* * *

We filter out all but the tiniest fraction of our sensory inputs; otherwise we would go mad. One by one, we label and dismiss these signals, unconsciously telling them, "You're not important; don't bother me."

> "Novelty itself will always rivet one's attention. There is that unique moment when one confronts something new and astonishment begins. Whatever it is, it looms brightly, its edges sharp, its details ravishing, in a hard clear light; just beholding it is a form of revelation, a new sensory litany. But the second time one sees it, the mind says, Oh, that again, another wing walker, another moon landing. And soon, when it's become commonplace, the brain begins slurring the details, recognizing it too quickly, by just a few of its features." [Ackerman, 1990]

Schema modification begins immediately after first exposure, with a memory replay of the incident and the attempt to name it and associate it with other incidents. Schema development may be immediate or gradual, and this rate is affected by many factors, particularly emotion and pain. Consider the following three experiments:

> A slug is fed a novel food, and then it is injected with a chemical that causes it to regurgitate. From only this one learning experience, it will immediately react to every future taste of this food by regurgitating. [Calvin, 1986]

> A rat is trained to press a lever for food. If it receives food every time that it presses the lever, it will learn much faster than if it is only intermittently reinforced.

> A rat is trained to press a lever for food. Randomly, however, it receives a shock rather than food when it presses the lever. If this is the only food source, the rat will continue to press the lever, but it is likely to develop irrational behavior (e.g., biting its handler) in other respects.

Emotional content and conflict enhance schema formation and memory:

> Consider two nearly identical physics lectures on ballistics. The only difference is that the instructor begins one by silently loading a gun and placing it on the lectern so that it is visible to the students throughout the lecture. Which class will remember ballistics better?

A schema may include non-relevant aspects, particularly those neural patterns that happened to be flowing during the first experience of the schema. Thus is superstition born, and in Chapter 4 we discussed the associated inductive fallacy of 'hasty generalization.' Non-relevant aspects, once established in the schema, are slow to disappear if one is not consciously aware of them. For example, a primary aim of psychotherapy is identification of behavior patterns formed in childhood and no longer appropriate. Deliberate schema modification can backfire:

> I once decided to add some enjoyment to the chore of dish washing by playing my favorite record whenever I washed dishes. It helped for a while, but soon I found that listening to that music without washing dishes seemed more like work than like recreation. I had created an associative bridge between the two sets of neural patterns, so that each triggered the emotional associations of the other.

Identification of a pattern does not require an exact match with the schema. Because we are identifying the schema holistically rather than by listing its components, we may not notice that some components are missing. Nor are extra, noncharacteristic components always detected. For example, people tend to overlook aspects of their own lives that are inconsistent with their current self-image [Goleman, 1992b]. The conditioned schema of previous experiences adds the missing component or ignores the superfluous component.

Memory can replay one schema or a series of schemata without any external cues, simply by activating the relevant neural pathways. Memory recollects the holistic schema that was triggered by the original experience; it does not replay the actual external events. Missing or overlooked elements of the schema are ignored in the memories. Thus eyewitness testimony is notoriously poor. Thus scientists trust their written records much more than their recollections of an experiment. Neural pathways are reinforced by any repeat flow -- real or recalled. Each recollection has the potential of modifying a memory. I may think that I am remembering an incident from early childhood, but more likely I am actually recalling an often-repeated recollection rather than the initial experience. Recollection can be colored by associated desire. Some individuals are particularly prone to self-serving memory, but everyone is affected to some degree. Thus, when I described the first football game near the start of this chapter, I said:

> "I prefer to think that our team was inspired by this event, that we stopped the opponents' advance just short of a touchdown, and that we recovered the ball and drove for the winning touchdown. Indeed, I do vaguely remember that happening, but I am not certain."

The individual always attempts schema identification based on available cues, whether or not those cues are sufficient for unique identification. In such cases the subconscious supplies the answer that seems most plausible in the current environment. Ittelson and Kilpatrick [1951] argue persuasively that optical illusions are simply a readily investigated aspect of the much broader phenomenon of subconsciously probabilistic schema identification.

> "Resulting perceptions are not absolute revelations of 'what is out there' but are in the nature of probabilities or predictions based on past experience. These predictions are not always reliable, as the [optical illusion] demonstrations make clear." [Ittelson and Kilpatrick, 1951]

> "Perception is not determined simply by the stimulus patterns; rather it is a dynamic searching for the best interpretation of the available data. . . Perception involves going beyond the immediately given evidence of the senses: this evidence is assessed on many grounds and generally we make the best bet. . . Indeed, we may say that a perceived object *is* a hypothesis, suggested and tested by sensory data." [Gregory, 1966]

* * *

No wonder expectations affect our observations! With a perceptual system geared to understanding the present by matching it to previous experience, of course we are prone to see what we expect to see, overlook anomalies, and find that evidence confirms our beliefs (Figure 22). Remarkably, the scientist's propensity for identifying familiar patterns is combined with a hunger for discovering new patterns. For the fallible internal photographer, it doesn't matter whether the spectacle is the expansion of the universe or the fate of a bug:

> "So much of the fascinating insect activity around escapes me. No matter how much I see, I miss far more. Under my feet, in front of my eyes, at my very finger's end, events are transpiring of fascinating interest, if I but knew enough or was fortunate enough to see them. If the world is dull, it is because we are blind and deaf and dumb; because we know too little to sense the drama around us." [Fabre, cited by Teale, 1959]

* * *

Postmodernism

> "Science is a social phenomenon. . . It progresses by hunch, vision, and intuition. Much of its change through time is not a closer approach to absolute truth, but the alteration of cultural contexts that influence it. Facts are not pure information; culture also influences what we see and how we see it. Theories are not inexorable deductions from facts; most rely on imagination, which is cultural." [Gould, 1981]

In the mid-twentieth century, the arts were dominated by modernism, which emphasized form and technique. Many people thought modernism was excessively restrictive. In reaction, the postmodern movement was born in the 1960's, embracing a freedom and diversity of styles. Postmodern thinking abandoned the primacy of linear, goal-oriented behavior and adopted a more empathetic, multipath approach that valued diverse ethnic and cultural perspectives. Postmodernism encompassed the social movements of religious and ethnic groups, feminists, and gays, promoting pluralism of personal realities.

According to the postmodern critique, objective truth is a dangerous illusion, developed by a cultural 'elite' but sold as a valid multicultural description. Cultural influences are so pervasive that truth, definite knowledge, and objectivity are unobtainable. Consequently, we should qualify all

findings by specifying the investigator's cultural framework, and we should encourage development of multiple alternative perspectives (e.g., feminist, African American, non-Western).

During the last two decades, postmodernism has become the dominant movement of literature, art, philosophy, and history. It has also shaken up some of the social sciences, especially anthropology and sociology. It is, at the moment, nearly unknown among physical scientists. However, some proponents of postmodernism claim that it applies to all sciences. Most postmodernists distrust claims of universality and definite knowledge, and some therefore distrust the goals and products of science.

> "The mythology of science asserts that with many different scientists all asking their own questions and evaluating the answers independently, whatever personal bias creeps into their individual answers is cancelled out when the large picture is put together. . . But since, in fact, they have been predominantly university-trained white males from privileged social backgrounds, the bias has been narrow and the product often reveals more about the investigator than about the subject being researched." [Hubbard, 1979]

Postmodern literary criticism seeks deconstruction of the cultural and social context of literary works. More broadly, deconstruction analysis is thought to be appropriate for any claim of knowledge, including those of scientists. For example, postmodern anthropologists recognize that many previous interpretations were based on overlaying twentieth-century WASP perspectives onto cultures with totally different worldviews. They reject the quest for universal anthropological generalizations and laws, instead emphasizing the local perspective of a society and of groups within that society [Thomas, 1998].

The major issues for the sciences are those introduced earlier in this chapter. Theories and concepts inspire data collection, determining what kinds of observations are considered to be worthwhile. Resulting observations are theory-laden, in the sense of being inseparable from the theories, concepts, values, and assumptions associated with them. Many values and assumptions are universal (e.g., space and time) and some are nearly so (e.g., causality) and therefore reasonably safe. If, however, some values and assumptions are cultural rather than universal, then associated scientific results are cultural rather than universal. Other scientists with different backgrounds might reach incompatible conclusions that are equally valid.

> Bertrand Russell [1927] commented wryly on how a group's philosophical approach can affect its experimental results:"Animals studied by Americans rush about frantically, with an incredible display of hustle and pep, and at last achieve the desired result by chance. Animals observed by Germans sit still and think, and at last evolve the solution out of their inner consciousness."

The postmodern critique challenges us to consider the extent to which social constructions may bias our scientific values, assumptions, and even the overall conceptual framework of our own scientific discipline. Many hypotheses and experiments are legitimately subject to more than one interpretation; would a different culture or ethnic group have reached the same conclusion?

Most scientists are likely to conclude that their scientific discipline is robust enough to be unshaken by the postmodern critique. True, observations are theory-laden and concepts are value-laden, but the most critical data and hypotheses are powerful enough to demonstrate explanatory power with greater scope than their origins. Politics and the economics of funding availability undoubtedly affect the pace of progress in the various sciences, but rarely their conclusions. Social influences such as a power elite are capable at most of a temporary disruption of the scientific progress of a science.

> In the first section of this chapter, Jarrard's case-study examples include two football games and a murder. In the previous chapter, Jarrard uses three military

quotes, a naval example, an analogy to military strategy and tactics, and two competitive-chess quotes. Clearly, Jarrard is an American male.

> Wilford [1992a] offers a disturbing insight into a scientific field that today is questioning its fundamentals. The discipline is anthropology, and many anthropologists wonder how much of the field will survive this self analysis unscathed. The trigger was an apparently innocuous discovery about the Maori legend of colonization of New Zealand, a legend that describes an heroic long-distance migration in seven great canoes. Much of the legend, anthropologists now think, arose from the imaginative interpretation of anthropologists. Yet significantly, the Maoris now believe the legend as part of their culture.
>
> Can anthropologists hope to achieve an objective knowledge of any culture, if that culture's perceptive and analytical processes are inescapably molded by a different culture? Are they, in seeking to describe a tradition, actually inventing and sometimes imposing one? If cultural traditions are continuously evolving due to internal and external forces, where does the scientist seek objective reality?

Are the anthropologists alone in their plight?

* * *

Pitfalls of Subjectivity

> "The nature of scientific method is such that one must suppress one's hopes and wishes, and at some stages even one's intuition. In fact the distrust of self takes the form of setting traps to expose one's own fallacies." [Baker, 1970]

How can we reconcile the profound success of science with the conclusion that the perception process makes objectivity an unobtainable ideal? Apparently, science depends less on complete objectivity than most of us imagine. Perhaps we do use a biased balance to weigh and evaluate data. All balances are biased, but those who are aware of the limitations can use them effectively. To improve the accuracy of a balance, we must know its sources of error.

Pitfalls of subjectivity abound. We can find them in experimental designs, execution of experiments, data interpretations, and publications. Some can be avoided entirely; some can only be reduced.

Experimental Design

• **ignoring relevant variables**: Some variables are ignored because of sloppiness. For example, many experimental designs ignore instrument drift, even though its bias can be removed. Often, however, intentions are commendable but psychology intervenes.

1) We tend to ignore those variables that we consider irrelevant, even if other scientists have suggested that these variables are significant.

2) We ignore variables if we know of no way to remove them, because considering them forces us to admit that the experiment has ambiguities.

3) If two variables may be responsible for an effect, we concentrate on the dominant one and ignore the other.

4) If the influence of a dominant variable must be removed, we are likely to ignore ways of removing it completely. We unconsciously let it exert at least residual effects [Kuhn et al., 1988].

• **confirmation bias**:

1) During the literature review that precedes experiment, we may preferentially seek and find evidence that confirms our beliefs or preferred hypothesis.

2) We select the experiment most likely to support our beliefs. This insidiously frequent pitfall allows us to maintain the illusion of objectivity (for us as well as for others) by carrying out a rigorous experiment, while nevertheless obtaining a result that is comfortably consistent with expectations and desires.

This approach can hurt the individual more than the scientific community. When two conflicting schools of thought each generate supporting information, the battling individuals simply grow more polarized, yet the community *may* weigh the conflicting evidence more objectively. Individuals seeking to confirm their hypothesis may overlook ways of refuting it, but a skeptical scientific community is less likely to make that mistake.

• **biased sampling**: Subjective sampling that unconsciously favors the desired outcome is easily avoided by randomization. Too often, however, we fail to consider the relevance of this problem during experimental design, when countermeasures are still available.

• **wish-fulfilling assumption**: In conceiving an experiment, we may realize that it could be valuable and diagnostic *if* a certain assumption were valid or *if* a certain variable could be controlled. Strong desire for an obstacle to disappear tempts us to conclude that it is not really an obstacle.

Experiment Execution

• **biased abortive measurements**: Sometimes a routine measurement may be aborted. Such data are rejected because of our subjective decision that a distraction or an intrusion by an uncontrolled variable has adversely affected that measurement's reliability. If we are monitoring the measurement results, then our data expectations can influence the decision to abort or continue a measurement. Aborted measurements are seldom mentioned in publications, because they weaken reader confidence in the experiment (and maybe even in the experimenter).

The biasing effect can be reduced in several ways: (1) 'blind' measurements, during which we are unaware of the data's consistency or inconsistency with the tested hypothesis; (2) *a priori* selection of criteria for aborting a measurement; and (3) completion of all measurements, followed by discussion in the publication of the rationale for rejecting some.

• **biased rejection of measurements**: Unanticipated factors and uncontrolled variables can intrude on an experiment, potentially affecting the reliability of associated data. Data rejection is one solution. Many data-rejection decisions are influenced by expectations concerning what the data 'should be'.

Rejection may occur as soon as the measurement is completed or in the analysis stage. As with aborted data, rejected measurements should be, but seldom are, mentioned in publication. Data-rejection bias is avoidable, with the same precautions as those listed above for reducing bias from aborted measurements.

• **biased mistakes**: People make mistakes, and elaborate error checking can reduce but not totally eliminate mistaken observations. Particularly in the field of parapsychology where subtle statistical

effects are being detected (e.g., card guessing tests for extrasensory perception, or ESP), much research has investigated the phenomenon of 'motivated scoring errors.' Scoring hits or misses on such tests appears to be objective: either the guess matched the card or it did not. Mistakes, however, are more subjective and biased: believers in ESP tend to record a miss as a hit, and nonbelievers tend to score hits as misses.

Parapsychology experimenters long ago adapted experimental design by creating blinds to prevent motivated scoring errors, but researchers in most other fields are unaware of or unworried by the problem. Motivated scoring errors are subconscious, not deliberate. Most scientists would be offended by the suggestion that they were vulnerable to such mistakes, but you and I have made and will make the following subconsciously biasing mistakes:

1) errors in matching empirical results to predictions,

2) errors in listing and copying results,

3) accidental omissions of data, and

4) mistakes in calculations.

• **missing the unexpected**: Even 'obvious' features can be missed if they are unexpected. The flash-card experiment, discussed earlier in this chapter, was a memorable example of this pitfall. Unexpected results can be superior to expected ones: they can lead to insight and discovery of major new phenomena (Chapter 8). Some common oversights are: (1) failing to notice disparate results among a mass of familiar results; (2) seeing but rationalizing unexpected results; and (3) recording but failing to follow-up or publish unexpected results.

• **biased checking of results**: To avoid mistakes, we normally check some calculations and experimental results. To the extent that it is feasible, we try to check all calculations and tabulations, but in practice we cannot repeat every step. Many researchers selectively check only those results that are anomalous in some way; such data presumably are more likely to contain an error than are results that look OK. The reasoning is valid, but we must recognize that this biased checking imparts a tendency to obtain results that fulfill expectations. If we perform many experiments and seldom make mistakes, the bias is minor. For a complex set of calculations that could be affected substantially by a single mistake, however, we must beware the tendency to let the final answer influence the decision whether or not to check the calculations. Biased checking of results is closely related to the two preceding pitfalls of making motivated mistakes and missing the unexpected.

• **missing important 'background' characteristics**: Experiments can be affected by a bias of human senses, which are more sensitive to detecting change than to noticing constant detail [Beveridge, 1955]. In the midst of adjusting an independent variable and recording responses of a dependent variable, it is easy to miss subtle changes in yet another variable or to miss a constant source of bias. Exploitation of this pitfall is the key to many magicians' tricks; they call it misdirection. Our concern is not misdirection but perception bias. Einstein [1879-1955] said, "Raffiniert is der Herrgott, aber boshaft ist er nicht" ("God is subtle, but he is not malicious"), meaning that nature's secrets are concealed through subtlety rather than trickery.

• **placebo effect**: When human subjects are involved (e.g., psychology, sociology, and some biology experiments), their responses can reflect their expectations. For example, if given a placebo (a pill containing no medicine), some subjects in medical experiments show a real, measurable improvement in their medical problems due to their expectation that this 'medicine' is beneficial.

Many scientists avoid consideration of mind/body interactions; scientific recognition of the placebo effect is an exception. This pitfall is familiar to nearly all scientists who use human subjects. It is avoidable through the use of a *blind*: the experimenter who interacts with the subject does not know whether the subject is receiving medicine or placebo.

• **subconscious signaling**: We can influence an experimental subject's response involuntarily, through subconsciously signaling. As with the placebo effect, this pitfall is avoidable through the use of blinds.

Data Interpretation

• **confirmation bias in data interpretation**: Data interpretation is subjective, and it can be dominated by prior belief. We should separate the interpretation of new data from the comparison of these data to prior results. Most publications do attempt to distinguish data interpretation from reconciliation with previous results. Often, however, the boundary is fuzzy, and we bias the immediate data interpretation in favor of our expectations from previous data.

• **hidden control of prior theories on conclusions**: Ideally, we should compare old and new data face-to-face, but too often we simply recall the conclusions based on previous experiments. Consequently, we may not realize how little chance we are giving a new result to displace our prior theories and conclusions. This problem is considered in more detail in that part of the next chapter devoted to paradigms.

• **biased evaluation of subjective data**: Prior theories always influence our evaluation of subjective data, even if we are alert to this bias and try to be objective. We can avoid this pitfall through an experimental method that uses a blind: the person rating the subjective data does not know whether the data are from a control or test group, or what the relationship is of each datum to the variable of interest. However, researchers in most disciplines never even think of using a blind; nor can we use a blind when evaluating published studies by others.

• **changing standards of interpretation**: Subjectivity permits us to change standards within a dataset or between datasets, to exclude data that are inconsistent with our prior beliefs while including data that are more dubious but consistent with our expectations [Gould, 1981]. A similar phenomenon is the overestimation of correlation quality when one expects a correlation and underestimation of correlation quality when no correlation is expected [Kuhn et al., 1988].

Publication

• **language bias**: We may use different words to describe the same experimental result, to minimize or maximize its importance (e.g., 'somewhat larger' vs. 'substantially larger'). Sarcasm and ridicule should have no place in a scientific article; they undermine data or interpretations in a manner that obscures the actual strengths and weaknesses of evidence.

• **advocacy masquerading as objectivity**: We may appear to be objective in our interpretations, while actually letting them be strongly influenced by prior theories. Gould [1981], who invented this expression, both criticizes and falls victim to this pitfall.

• **weighting one's own data preferentially**: This problem is universal. We know the strengths and weaknesses of personally-obtained data much better than those of other published results. Or so we rationalize our preference for our own data. Yet both ego and self-esteem play a role in the frequent subjective decision to state in print that one's own evidence supersedes conflicting evidence of others.

• **failure to publish negative results**: Many experimental results are never published. Perhaps the results are humdrum or the experimental design is flawed, but often we fail to publish simply because the results are negative: we do not understand them, they fail to produce a predicted pattern, or they are otherwise inconsistent with expectations. If we submit negative results for publication, the manuscript is likely to be rejected because of unfavorable reviews ('not significant'). I have even heard of a journal deliberately introducing this bias by announcing that they will not accept negative results for publication. Yet a diagnostic demonstration of negative results can be extremely useful -- it can force us to change our theories.

• **concealing the pitfalls above**: The myth of objectivity usually compels us to conceal evidence that our experiment is subject to any of the pitfalls above. Perhaps we make a conscious decision not to bog down the publication with subjective ambiguities. More likely, we are unaware or only peripherally cognizant of the pitfalls. Social scientists recognize the difficulty of completely avoiding influence of the researcher's values on a result. Therefore they often use a twofold approach: try to minimize bias, and also specifically spell out one's values in the publication, so that the reader can judge success.

* * *

> "The great investigator is primarily and preeminently the man who is rich in hypotheses. In the plenitude of his wealth he can spare the weaklings without regret; and having many from which to select, his mind maintains a judicial attitude. The man who can produce but one, cherishes and champions that one as his own, and is blind to its faults. With such men, the testing of alternative hypotheses is accomplished only through controversy. Crucial observations are warped by prejudice, and the triumph of the truth is delayed." [Gilbert, 1886]

* * *

Pitfall Examples

> Penzias and Wilson [1965] discovered the background radiation of the universe by accident. When their horn antenna detected this signal that was inconsistent with prevailing theories, their first reaction was that their instrument somehow was generating noise. They cleaned it, dismantled it, changed out parts, but they were still unable to prevent their instrument from detecting this apparent background radiation. Finally they were forced to conclude that they had discovered a real effect.
>
> *Pitfalls: biased checking of results;*
> *biased rejection of measurements;*
>
> Throughout the 20th century, scientists from many countries have sought techniques for successfully predicting earthquakes. In 1900 Imamura predicted that a major earthquake would hit Tokyo, and for two decades he campaigned unsuccessfully to persuade people to prepare. In 1923, 160,000 people died in the Tokyo

earthquake. Many predictions have been made since then, by various scientists, based on diverse techniques. Still we lack reliable techniques for earthquake prediction. Said Lucy Jones [1990] of the U.S. Geological Survey, "When people want something too much, it's very easy to overestimate what you've got." Most of the altruistic predictions suffered from one of the following *pitfalls:*

wish-fulfilling assumption or treatment of a variable;
biased sampling; or
confirmation bias in data interpretation.

* * *

The following examples were used by Gould [1981] to illustrate the severe societal damage that lapses in scientific objectivity can inflict.

If an objective, quantitative measure of intelligence quotient (IQ), independent of environment, could be found, then education and training possibly could be optimized by tailoring them to this innate ability. This rationale was responsible for development of the Army Mental Tests, which were used on World War I draftees. Among the results of these tests were the observations that white immigrants scored lower than white native-born subjects, and immigrant scores showed a strong correlation with the number of years since immigration. The obvious explanation for these observations is that the tests retained some cultural and language biases. The actual interpretation, which was controlled by desire for the tests to be objective measures of IQ, was the following: a combination of lower intelligence in Europeans than in Americans and of declining intelligence of immigrants. This faulty reasoning was used in establishing the 1924 Immigration Restriction Act. [Gould, 1981]

Pitfalls: wish-fulfilling assumption or treatment of variable;
ignoring relevant variables;
hidden control of prior theories on conclusions.

Bean 'proved' black inferiority by measuring brain volumes of blacks and whites and demonstrating statistically that black brains are smaller than white brains. His mentor Mall replicated the experiment, however, and found no significant difference in average brain size. The discrepancy of results is attributable to Mall's use of a blind: at the time of measurement, he had no clues as to whether the brain he was measuring came from a black or white person. Bean's many measurements had simply reflected his expectations. [Gould, 1981]

Pitfalls: biased evaluation of subjective data;
advocacy masquerading as objectivity.

In order to demonstrate that blacks are more closely related to apes than whites are, Paul Broca examined a wide variety of anatomical characteristics, found those showing the desired correlation, and then made a large number of careful and reliable measurements of only those characteristics. [Gould, 1981]

Pitfalls: confirmation bias in experimental design
(selecting the experiment most likely to support one's beliefs);
confirmation bias in data interpretation;
hidden control or prior theories on conclusions;
advocacy masquerading as objectivity.

* * *

Group Objectivity

> "The objectivity of science is not a matter of the individual scientists but rather the social result of their mutual criticism." [Popper, 1976]

> "Creating a scenario may be best done inside a single head; trying to find exceptions to the scenario is surely best done by many heads." [Calvin, 1986]

One can reduce deliberately the influence of the pitfalls above on one's research. One cannot eliminate the effects of personal involvement and personal opinions, nor is it desirable to do so. *Reliability of conclusions, not objectivity, is the final goal. Objectivity simply assists us in obtaining an accurate conclusion*. Subjectivity is essential to the advance of science, because scientific conclusions are seldom purely deductive; usually they must be evaluated subjectively in the light of other knowledge.

But what of experimenter bias? Given the success of science in spite of such partiality, can it actually be a positive force in science? Or does science have internal checks and balances to reduce the adverse effects of bias? Several philosophers of science [e.g., Popper, 1976; Mannoia, 1980; Boyd, 1985] argue that a scientific community can make objective consensus decisions in spite of the biases of individual proponents.

It's said [e.g., Beveridge, 1955] that only the creator of a hypothesis *believes* it (others are dubious), yet only the experimenter *doubts* his experiment (others cannot know all of the experimental uncertainties). Of course, group evaluation is much less credible and unanimous than this generalization implies. The point, instead, is that the individual scientist and the scientific community have markedly different perspectives.

Replication is one key to the power of group objectivity. Replicatability is expected for all experimental results: it should be possible for other scientists to repeat the experiment and obtain similar results. Published descriptions of experimental technique need to be complete enough to permit that replication. As discussed in Chapter 2, follow-up studies by other investigators usually go beyond the original, often by increasing precision or by isolating variables. Exact replication of the initial experiment seldom is attempted, unless those original experimental results conflict with prior concepts or with later experiments. Individual lapses of objectivity are likely to be detected by the variety of perspectives, assumptions, and experimental techniques employed by the scientific community.

Perhaps science is much more objective than individual scientists, in the same way that American politics is more objective than either of the two political parties or the individual politicians within those parties. Politicians are infamous for harboring bias toward their own special interests, yet no doubt they seek benefits for their constituents more often than they pursue personal power or glory. Indeed, concern with personal power or glory is more relevant to scientists than we like to admit (Chapter 9).

The strength of the political party system is that two conflicting views are advocated by two groups, each trying to explain the premises, logic, and strengths of one perspective and the weaknesses of the other point of view. *If one pays attention to both viewpoints*, then hopefully one has all of the information needed for a reliable decision. Conflicting evidence confronts people with the

necessity of personally evaluating evidence. Unfortunately, both viewpoints are not fully developed in the same place; we must actively seek the conflicting arguments.

Suppose a scientist writes a paper that has repeated ambivalent statements such as "X may be true, as indicated by Y and Z; on the other hand, . . ." That scientist may be objective, but the paper probably has little impact, because it leaves most readers with the impression that the subject is a morass of conflicting, irreconcilable evidence. Only a few readers will take the time to evaluate the conflicting evidence.

Now suppose that one scientist writes a paper saying, "X, not Y, is probable because. . ." and another scientist counters with "Y, not X, is probable because. . ." Clearly, the reader is challenged to evaluate these viewpoints and reach a personal conclusion. This dynamic opposition may generate a healthy and active debate plus subsequent research. "All things come into being and pass away through strife" [Heraclitus, ~550-475 B.C.]. Science gains, and the only losers are the advocates of the minority view. Even they lose little prestige, because their role is remembered more for its active involvement in a fascinating problem than for being 'wrong', if the losers show in print that they have changed their minds because of more convincing evidence. In contrast, the loser who continues as a voice in the wilderness does lose credibility (even if he or she is right).

> "It is not enough to observe, experiment, theorize, calculate and communicate; we must also argue, criticize, debate, expound, summarize, and otherwise transform the information that we have obtained individually into reliable, well established, public knowledge." [Ziman, 1969]

Given two contradictory datasets or theories (e.g., light as waves vs. particles), the scientific community gains if some scientists simply assume each and then pursue its ramifications. This incremental work eventually may offer a reconciliation or solution of the original conflict. Temporary abandonment of objectivity thus can promote progress.

Science is not democratic. Often a lone dissenter sways the opinions of the scientific community. The only compulsion to follow the majority view is peer pressure, which we first discovered in elementary school and which haunts us the rest of our lives.

A consensus evolves from conflicting individual views most readily if the debating scientists have similar backgrounds. Divergent scientific backgrounds cause divergent expectations, substantially delaying evolution of a consensus. For example, geology has gone through prolonged polarizations of views between Northern Hemisphere and Southern Hemisphere geologists on the questions of continental drift and the origin of granites. In both cases, the locally observable geologic examples were more convincing to a community than were arguments based on geographically remote examples.

This heterogeneity of perspectives and objectives is an asset to science, in spite of delayed consensus. It promotes group objectivity and improves error-checking of ideas. In contrast, groups that are isolated, homogeneous, or hierarchical tend to have similar perspectives. For example, Soviet science has lagged Western science in several fields, due partly to isolation and partly to a hierarchy that discouraged challenging of the leaders' opinions.

* * *

> The cold-fusion fiasco is an excellent example of the robustness of group objectivity, in contrast to individual subjectivity. In 1989 Stanley Pons and Martin Fleischmann announced that they had produced nuclear fusion in a test tube under ordinary laboratory conditions. The announcement was premature: they had not rigorously isolated variables and thoroughly explored the phenomenon. The rush to public announcement, which did not even wait for simultaneous presentation to peers at a scientific meeting, was generated by several factors: the staggering possible benefit to humanity of cheap nuclear power, the Nobel-level accolades that would accrue to the

discoverers, the fear that another group might scoop them, executive decisions by the man who *was* president of my university, and exuberance.

The announcement ignited a fire-storm of attempts to replicate the experiments. Very few of those attempts succeeded. Pons and Fleischmann were accused of multiple lapses of objectivity: wish-fulfilling assumptions, confirmation bias, ignoring relevant variables, mistakes, missing important background characteristics, and optimistic interpretation.

In a remarkably short time, the scientific community had explored and discredited cold fusion. Group objectivity had triumphed: fellow scientists had shown that they were willing to entertain a theoretically ridiculous hypothesis and subject it to a suite of experimental tests.

Groups can, of course, temporarily succumb to the same objectivity lapses as individuals. N rays are an example.

Not long after Roentgen's discovery of X rays, Rene Blondlot published a related discovery: N rays, generated in substances such as heated metals and gases, refracted by aluminum prisms, and observed by phosphorescent detectors. If one sees what one expects to see, sometimes many can do the same. The enthusiastic exploration of N rays quickly led to dozens of publications on their 'observed' properties. Eventually, of course, failures to replicate led to more rigorous experiments and then to abandonment of the concept of N rays. The scientific community moved on, but Blondlot died still believing in his discovery.

We began this section with a paradox: how can it be possible for many subjective scientists to achieve objective knowledge? We concluded that science does have checks and balances that permit it to be much more objective than the individual scientists. The process is imperfect: groups are temporarily subject to the same subjectivity as individuals. Group 'objectivity' also has its own pitfalls. We shall postpone consideration of those pitfalls until the next chapter, however, so that we can see them from the perspective of Thomas Kuhn's remarkable insights into scientific paradigm.

Chapter 7: Evidence Evaluation and Scientific Progress

Scientists and philosophers of science share a concern for evidence evaluation and scientific progress. Their goals, however, are quite different. The philosophers find the process of science intrinsically interesting. Most are not trying to 'straighten out' the scientists and tell them how science should be done. Some of their conclusions do have possible implications for future scientific methods, but scientists seldom listen. Perhaps scientists' reactions are somewhat analogous to those of creative writers toward literary critics and academic literary analysts: often the doer is unappreciative of the outside reviewer.

Each scientist unconsciously selects criteria for evaluating hypotheses. Yet clearly it would be both confusing and professionally hazardous to adopt substantially different criteria than those used by one's peers. Judgment, not irrefutable evidence, is a foundation of science. Judgments that observations confirm or refute hypotheses are based on personal values: accuracy, simplicity, consistency, scope, progressiveness, utility, and expediency.

[Larson, 1987]

Different types of laws, or hypotheses, require different evaluation criteria [Carnap, 1966]. A **universal law** such as 'all ravens are black' is best tested by seeking a single exception. In contrast, a **statistical law** such as 'almost all ravens are black' or '99% of ravens are black' requires a statistical test that compares observed frequencies to hypothesized frequencies. Theoretical and empirical hypotheses call for contrasting evaluation techniques and standards. For example, a theoretical-physics hypothesis may concern properties that are not directly measurable and that must be inferred indirectly, and it may be judged more on simplicity and scope than on accuracy of fit to observations.

This chapter considers all of these aspects of evidence evaluation.

* * *

Critical thinking skills and mistakes begun in childhood survive the transition to adult. Naive conceptions do not simply disappear when a more mature thinking skill is developed; they must be consciously recognized as wrong and deliberately replaced. For example, children learn about causality first by treating all predecessors as causal ("if I wear a raincoat and avoid getting wet, I won't get a cold"). Only later and rather haphazardly is this superstitious approach supplanted by the skill of isolation of variables.

The most important childhood development in reasoning skill is obtaining conscious control over the interaction between theory and evidence [Kuhn et al., 1988]. Immature thinking fails to distinguish between theory and evidence. Scientific thinking requires critical evaluation of observations and of their impact on the validity of hypotheses. This skill is polished by practice -- particularly by coping with contradictory evidence and contradictory hypotheses.

In order to relate evidence to hypotheses effectively, the researcher needs three related skills [Kuhn et al., 1988]:

- The evidence must be analyzed independently of the hypothesis, *before* evaluating the relationship between data and hypothesis.
- One must be able to think *about* a hypothesis rather than just *with* it. If one allows the hypothesis to guide interpretation of the evidence, objective evidence evaluation is impossible.
- While considering the impact of evidence on the hypothesis, one must be able to ignore personal opinion of the affected hypothesis. Favorable and unfavorable evidence must be given a chance to affect the final conclusion.

Kuhn et al. [1988] find that these three skills, which are absent in children and are developed gradually during middle adolescence and beyond, are still below optimum even in most adults.

Like most college students, I memorized facts and absorbed concepts, but I was seldom faced -- at least in class-work -- with the 'inefficient' task of personally evaluating evidence and deciding what to believe. Imagine my surprise when I went to graduate school, began reading the scientific literature, and discovered that even some ridiculous ideas have proponents. Textbook learning does not teach us the necessity of evaluating every conclusion personally -- regardless of how famous the writer is, regardless of how meager one's own experience is.

Effective evidence evaluation requires active critical thinking, not passive acceptance of someone else's conclusion. The reader of a publication must become a reviewer who judges the evidence for and against the writer's conclusion.

Effective evidence evaluation is far more comprehensive than discrimination of whether statements are correct. It also involves assessment of the scope and ambiguities of observations, generalizations, and deductions, as well as the recognition of implicit and explicit assumptions. Have all perspectives been considered? Is any conclusion warranted by the evidence?

The evaluation techniques of this chapter can aid in this wresting of control from subconscious feelings and toward rational decision-making.

* * *

Judgment Values

Evidence evaluation, like scientific research in general, involves not only technique but also style. One's judgment of an hypothesis or evidence set is more a product of subjective values than of objective weighting factors. Those values, like scientific research style, are based on personal taste.

Prediction of observations is perhaps the most compelling type of confirmation or refutation. As discussed later in this chapter, the confirmatory power of evidence depends on how surprising the prediction is. A rule of thumb is:

In forming a hypothesis, value minimum astonishment;
in testing hypothesis predictions, value maximum astonishment.

Thus hypotheses that are simple and, at least in hindsight, obvious are valued over convoluted ones. In contrast, the more unexpected and outlandish a prediction is, the more compelling it is if found to be correct. For example, Einstein was a master at forming theories that were based on the simplest of premises, yet yielded seemingly absurd but verifiably correct predictions.

Prediction is always valued over retrodiction, the ability of a hypothesis to account for data already known. This difference in values is because prediction constitutes an independent test of an idea, whereas existing data may have been incorporated in concept formation. For example, a polynomial may fit a set of time-series data excellently, yet generate bizarre predictions for regions outside the range of the input data. On shakier ground are retrodictions consisting of data that existed when the hypothesis was developed but of which the discoverer was unaware. The discoverer rightly considers them to be independent and successful predictions; the fact that the experiment preceded the hypothesis is irrelevant. The reviewer, however, cannot know whether or not the idea's author was indirectly influenced by these data.

Comparison of a hypothesis to existing data is the first step in its testing, but this evaluation could have a hidden bias. The experiments were not designed specifically to test this hypothesis, so one must subjectively select 'appropriate' experiments and interpret departures from ideal experimental design. Predictions, in contrast, minimize these problems.

* * *

All scientists accept that hypothesis generation is subjective, but most cling to the myth that their evaluations of evidence are objective. Yet in recent decades the illusion of totally rational decision-making has collided with the technical difficulty of developing artificial intelligence (AI) programs. The successes and failures of AI suggest the scope of the problem. AI achieved rapid success in medical diagnosis, where each of an enormous number of potential symptoms has established statistical implications for potential diagnoses. In contrast, AI has progressed surprisingly slowly, in spite of great effort, in duplicating human language. Apparently, the 'rules' of grammar and 'definitions' of words are fuzzier and more qualitative than we had thought.

AI undoubtedly will expand dramatically during the next two decades, but its start has been sluggish, probably because of the subjectivity implicit in much scientific decision-making. "Every individual choice between competing theories depends on a mixture of objective and subjective factors, or of shared and individual criteria" [Kuhn, 1977]. These scientific decisions involve the weighing of competing advantages that are really not comparable or weighable. And even if one could develop a set of such weighting factors, we would find that they differ among individuals.

To identify these subjective weighting factors used in evidence evaluation, Kuhn [1977] asked "What are the characteristics of a good theory?" He identified five: accuracy, consistency, scope, simplicity, and fruitfulness. I add two others: utility and expediency. These are the seven main values on which we base our judgments concerning confirmation or refutation of hypotheses.

Accuracy -- and especially quantitative accuracy -- is the king of scientific values. Accuracy is the closest of the seven to an objective and compelling criterion. Accuracy is the value that is most closely linked to explanatory ability and prediction; hypotheses must accord with observations. Indeed, 2500 years after Pythagoras' fantasy of a mathematical description of nature, quantitative ac-

curacy is now the standard of excellence in all sciences that are capable of pragmatically embracing it.

The value placed on quantitative accuracy extends beyond the judging of hypotheses; it can affect one's choice of scientific field. Highly quantitative sciences are not intrinsically superior to nonquantitative sciences; individual tastes are not comparable.

Simplicity is a value that is implicit to the scientist's objective of identifying patterns, rules, and functional similarity among unique individual events. Yet all hypotheses seek order amid apparent complexity, so how does one apply the criterion of simplicity? William of Occam, a 14th-century English philosopher, developed 'Occam's Razor' as a method of cutting to the truth of a matter: "The simplest answer is the one most likely to be correct." Also known as the maxim of parsimony, Occam's Razor is an imperfect rule of thumb, but often it does select correctly among hypotheses that attempt to account for the same observations. The 'simplest answer' is not necessarily the one most easily comprehended. Often it is the one with the fewest assumptions, rationalizations, and particularly special cases, or it is the most elegant idea.

> Sherlock Holmes countered the emphasis on simplicity by saying, "When all other contingencies fail, whatever remains, however improbable, must be the truth" [Doyle, 1917]. Yet when scientists resort to hypothesizing the improbable, they usually discover the actual truth later, among options that had been overlooked.

I still remember a sign that I saw on a restroom paper-towel dispenser twenty years ago: "Why use two when one will do?" The advice is in accord with Occam's Razor: two or more hypotheses, each of which explains part of the observations, are less likely to be correct that one umbrella hypothesis that accounts for all of the data. Similarly, if an explanation becomes more and more complex as it is modified to account for incompatible observations, it becomes more suspect according to Occam's Razor.

Complexity can result, however, from the interactions among two or more simple phenomena. For example, simple fractal geometric rules of repetition, when applied at different scales, can result in apparently complex patterns such as branching river systems and branching trees. Molecular biologists have long puzzled over how simple amino acids made the evolutionary leap to complex DNA; now these researchers are exploring the possibility that a few simple rules may be responsible [Gleick, 1992c].

The value on simplicity leads most scientists to be distrustful of coincidences. We recognize that they occur, but we suspect that most mask simple relationships.

> Not everyone values simplicity similarly. Georg Ohm, a mathematics professor in Cologne, proposed in 1827 that electrical current in a wire is simply proportional to the potential difference between the wire ends. His colleagues considered this idea to be simplistic, and he was forced to resign his position. Eventually, his hypothesis was accepted and he resumed his academic career -- this time as professor of experimental physics. Today Ohm's Law, which says that potential difference equals the product of current and resistance (in ohms), is the most useful equation in electricity.

Consistency, an aspect of simplicity, is valued in all sciences. The hypothesis should be consistent with relevant concepts that have already been accepted, or else it will face the formidable hurdle of either overthrowing the established wisdom or uneasily coexisting with incompatible hypotheses. Such coexistence is rare; one example is the physics concept of complementarity,

discussed later in this section. An explanation must also be self-consistent: for example, all hypotheses are wrong, including this one.

> In 320 B.C., Pytheas of Massilia sailed beyond the northernmost limits of the known world, to the land of Thule north of Britain. When he returned, he claimed that in Thule the midsummer sun did not set. His contemporaries called this observation preposterous.

Scope is another aspect of simplicity. A hypothesis that only accounts for the observations that inspired it has little value. In contrast, a hypothesis with a broad explanatory power inspires confidence through its ability to find order in formerly disparate types of observations. Scope is the antidote to Popper's [1963] criticism that many similar confirmations can only marginally increase confidence in a hypothesis. A hypothesis with broad scope tends to be more amenable to diversified testing.

Progressiveness, or fruitfulness, is a seldom discussed value. Kuhn [1977] says simply that " a theory should be fruitful of new research findings: It should, that is, disclose new phenomena or previously unnoted relationships among those already known." Most hypotheses seek to disclose previously unnoted relationships. Yet some are dead ends, sparking no further research except the confirmation or refutation of that specific conjecture. In contrast, progressive hypotheses are valued because of their exciting implications for a variety of new research directions. Even if a fruitful idea is later determined to be wrong, it can constructively steer future research efforts. Oliver [1991] thinks that the best criterion for the value of a scientific publication is the "impacts of the paper on the flow of science," the extent to which it changes what other scientists do.

> "A great discovery is a fact whose appearance in science gives rise to shining ideas, whose light dispels many obscurities and shows us new paths." [Bernard, 1865]

> "A great discovery is not a terminus, but an avenue leading to regions hitherto unknown. We climb to the top of the peak and find that it reveals to us another higher than any we have yet seen, and so it goes on." [Thomson, 1961]

Utility is not just a crucial value for applied scientists; it is a common concern of all scientists. We scan journals and focus almost exclusively on articles that may be of some utility to us. To results that are not personally useful, we apply the most lethal hypothesis-evaluation technique: we ignore them. Similarly, the depth of our evaluation depends on the perceived relevance and utility of the hypothesis. When choosing between two hypotheses, we normally select the more pragmatic and useful one. For example, a useful empirical equation is often preferred over a rigorous theoretical equation, if the latter includes several variables that we are unable to estimate.

Expediency is concern for what is immediately advantageous, and scientific expediency favors acceptance of promised solutions to worrisome problems. Scientific anxiety is created when a ruling theory is threatened, or indeed whenever a discipline is faced with an apparently irreconcilable conflict -- perhaps between two incompatible hypotheses or perhaps between a strong hypothesis and a compelling dataset. Any evidence or ancillary explanation that promises remedy for the anxiety is likely to be received favorably -- almost gratefully -- because of the expediency factor. Valu-

ing expediency can pose a pitfall, leading us beyond objective evidence evaluation and obscuring a broader question: how much do I want the idea to be confirmed, for other reasons?

* * *

Like all values, these seven are "effective guidance in the presence of conflict and equivocation" [Kuhn, 1977], not rigid criteria that dictate an unambiguous conclusion. "The criteria of choice . . . function not as rules, which determine choice, but as values, which influence it." Like all values, these differ among individuals. Thus the disagreements between scientists about a hypothesis do not imply that one has misinterpreted data or made an error. More likely, they employ different subjective weightings of conflicting evidence. In Chapter 6, I argued that such disagreements are actually scientifically healthy and that they are an efficient means for advancing science; group objectivity grows from individuals' subjectivity.

Scientific values differ between fields, and they may evolve within a field. For example, engineers and applied scientists emphasize the value of social utility as a key evaluation criterion, and they differ with physicists concerning the relative value of fruitfulness and simplicity. Kuhn [1977] notes that quantitative accuracy has become an evaluation criterion for different sciences at different times: it was achievable and valued by astronomy many centuries ago; it reached mechanics three centuries ago, chemistry two centuries ago, and biology in this century.

Like all human values and unlike rules, the scientific values are implicitly imprecise and often contradictory. For example, more complex hypotheses are usually more accurate than simple ones, and hypotheses with a narrow scope tend to be more accurate than those with a broad scope. Even a single value such as accuracy may have contradictory implications: a hypothesis may be more accurate than a competing idea in one respect and less accurate in another, and the scientist must decide which is more diagnostic.

> An extreme example of the conflict between values is the quantum mechanics concept of complementarity, which achieves utility and expediency by abandoning consistency. According to complementarity, no single theory can account for all aspects of quantum mechanics. Concepts such as light as waves and light as particles are complementary. Similarly, the concepts of determining position precisely and determining momentum precisely are complementary. Furthermore, the concept of determining location in space-time is complementary to the concept of determinacy. In each case the pair of concepts is apparently contradictory; assuming one seems to exclude the other in an individual experiment. But full 'explanation' requires both concepts to be embraced, each in different situations. Complementary concepts only seem to be contradictory, because our perceptions are unable to reconcile the contradictions. The actual physical universe, independent of our observing process, has no such contradictions.

* * *

Evaluation Aids

Scientific progress depends on proper appraisal of evidence, on successful rejection of incorrect hypotheses and adoption of correct (or at least useful) hypotheses. Yet the evaluation techniques employed most often are incredibly haphazard, leading to conclusions such as 'sounds reasonable' or 'seems rather dubious'.

Evaluation of evidence is a scientific skill, perhaps the most important ability of a successful scientist. Like any skill, its techniques must be practiced deliberately and systematically, before one

can trust its subconscious or casual use. For those who are mastering evidence evaluation, and even occasionally for experienced scientists, evaluation aids are useful. Here we describe three such techniques: model/observation tables, outlines, and concept maps.

* * *

A model/observation table succinctly compares several competing hypotheses. Usually various observations are relevant, with some favoring one idea while others favor another. Ideally, the scientist would examine each hypothesis systematically, reject those refuted by one or more evidence sets, and conclude that only a single hypothesis survives unscathed. In practice, we generally must weigh many inconclusive and partially contradictory data. The challenge to the scientist is to consider simultaneously this variety of evidence; a model/observation table is one way.

The model/observation table is a specialized and somewhat qualitative version of a truth table: list the models (or hypotheses) horizontally, list the relevant observations vertically, and then symbolically summarize the consistency of each observation with each model. Select symbols that are readily translatable into position along the continuum from strong confirmation to strong refutation:

+: strong confirmation
+: weak or ambiguous confirmation
0: not relevant, or no data available
(alternatively, use a blank if '0' implies 'no' to you)
-: weak or ambiguous refutation
--: strong refutation.

Table 11. Example of a model/observation table.

Observation	[A, '75]	[B&C, '76]	[D, '80]	[D&E, '81]
x/y correlation	+	-	+	+
y=3.7x	+	-	+	**+**
no y/z correlation	**--**	+	+	+
w= 5.2	0	0	+	**+**
x<w	+	+	+	+

For example, Table 11 summarizes the consistency of four published models with a group of five experimental findings. A quick scan of this table permits us to see that the leading hypotheses are those of D [1980] and of D & E [1981]; the latter is somewhat more successful but not decisively so. The hypothesis of A [1975], though consistent with many observations, is refuted by the observation of no y/z correlation. The hypothesis of B&C [1976] has mixed and unimpressive consistency with the observations. This quick overview allows identification of which observations are the most useful and consequently warrant the most careful attention. For example, the observation that x<w obviously is of no help in distinguishing among the possibilities.

The model/observation table is an easy way to focus one's attention onto the most diagnostic relationships among observations and hypotheses. It counteracts the universal tendency toward letting one relationship dominate one's thoughts. It encourages systematic evaluation of all relevant types of evidence. The table is not meant to be a simple tabulation of consistency scores, resulting

in a bottom line success/failure score for each hypothesis. It cannot be that quantitatively objective, for the various observations are of unequal reliability and significance, in ways not readily reducible to a +/- symbol. Nevertheless, even experienced scientists often are surprised at how effectively this underused technique can draw their attention to the crux of a problem.

An outline is a familiar technique that is readily adapted for evidence evaluation. An outline works effectively for analysis of one or two major hypotheses. For multiple hypotheses, it has considerable redundancy because different hypotheses are affected by the same arguments. In contrast, the model/observation table is more compact and is concentrated on identifying differences among hypotheses. Like the model/observation table, an outline permits both arguments for and arguments against a hypothesis. It also permits nested hypotheses: often the premise for one conclusion has its own premises, strengths, and weaknesses. An evidence-evaluation outline might look like the following:

I. Hypothesis

A) argument for hypothesis

1) primary confirmation of A

2) secondary confirmation of A

3) ambiguity

B) strong argument for hypothesis

1) primary confirmation of B

2) secondary confirmation of B

3) <u>But</u> evidence against B

a) confirmation of #3

b) <u>But</u> alternative explanation for #3

A less structured alternative to outlines and model/observation tables is the concept map, a flowchart that summarizes the known (or inferred) relationships among a suite of concepts. It is adaptable as a learning aid or as a method of evidence evaluation; at present it is used primarily as the former. Figure 23 illustrates the technique with a high-school-level concept map of sports [Arnaudin et al., 1984].

Concept mapping is based on a learning theory called cognitive association [Ausubel et al., 1978]. Cognitive association goes beyond the fixed patterns of simple memorization; like science, it evolves to encompass new knowledge. It employs the synergy of linking a new idea to existing ones: the new concept is easier to remember, and it subtly changes one's perceptions of previously known ones. Additional ideas are subsumed into the existing conceptual framework and, like analogies, gain meaning from familiarity of patterns.

Based on teaching concept mapping to several hundred students, Arnaudin et al. [1984] reach the following conclusions about this technique:

- it is an effective study technique.
- it improves one's ability to comprehend complex phenomena, by dissecting them into graspable components and links.

• it helps one to identify gaps in knowledge and understanding, thereby lending a goal-oriented aspect to further learning.

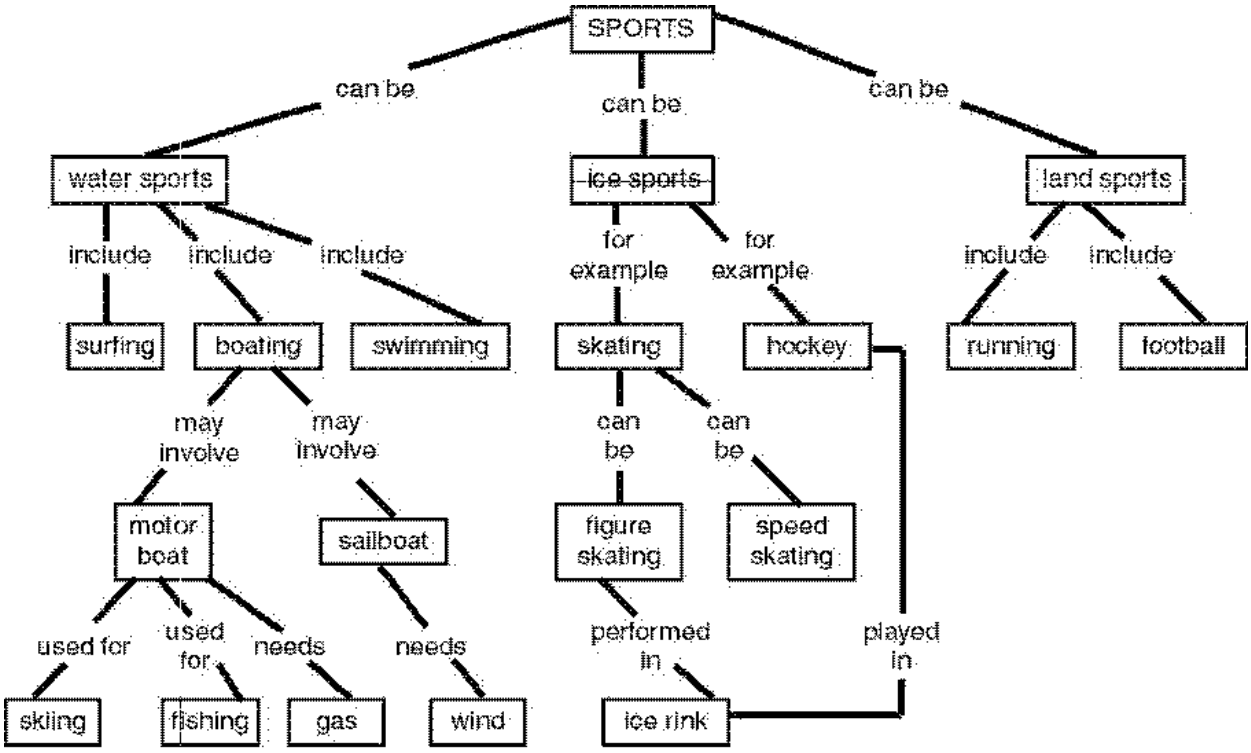


Figure 23. Example of a concept map of sports, demonstrating the use of concept maps for multi-level classifications [modified from Arnaudin et al., 1984].

A scientific publication can be concept mapped with the following seven-step procedure, adapted from one developed by J. D. Novak [Arnaudin et al., 1984]:

1) read the publication, highlighting or underlining key 'concepts' as you go. 'Concepts' can be hypotheses, assumptions, equations, or experiments, but not relationships.

2) skim back through the publication, systematically highlighting previously overlooked concepts that seem relevant to the overall context.

3) transfer all of the highlighted concepts to a list. Try to list the most general ones near the top and the less inclusive, more specific ones near the bottom. Sometimes an entire suite of related concepts can be encompassed in a larger-scale box representing a packaged general theory.

4) transfer the list onto a concept 'map', where broad categories are placed near the top of the map and successively more restrictive categories are placed successively lower on the map. Place similar concepts or categories on the same level, grouping related ones. Draw lines linking concepts on different levels. Label each line with a simple linking word that identifies the relationship between the pair of ideas.

5) 'branch out', adding concepts and links that were not in the publication but are suggested by examination of the map.

6) create 'cross-links', identifying connections between concepts that may be more distant than the simple downward branching of the overall concept map. This cross-linking procedure may even suggest a radical redrawing of the map, thereby simplifying its structure.

7) highlight, or weight, key concepts and links with bold lines or boxes, and use dashed lines and question marks for suspect portions.

Am I insulting the reader by suggesting that a high-school or college learning technique such as in Figure 23 is also useful for the professional scientist? Long before the term concept mapping was invented, a very similar flowchart technique was used by scientists and other professionals, for the identical purpose of visualizing the relationships among complex phenomena. For example, Figure 24 [Bronowski, 1973] is a page from the notes of John von Neumann, one of the most outstanding mathematicians of the 20th century; apparently his photographic memory did not preclude the usefulness of conceptual flowcharts. Figures 4 and 22 are additional examples. For the scientist who is analyzing and evaluating a scientific article, or who is trying to work through a complex idea, concept mapping can be a visualization and evaluation aid.

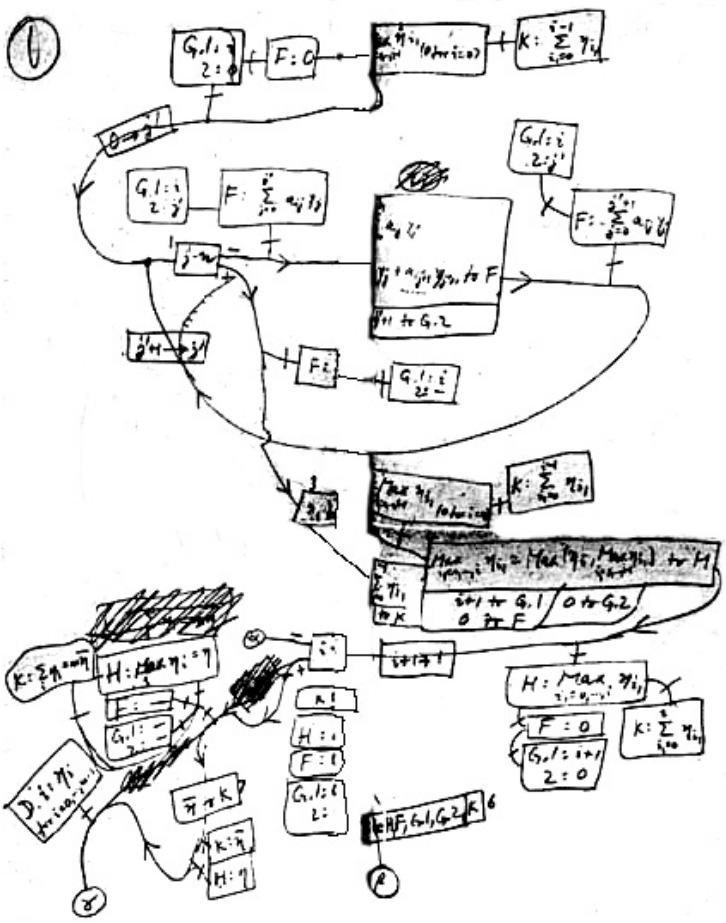


Figure 24. Concept map drawn by John von Neumann[Bronowski, 1973]

* * *

Model/observation tables, outlines, and concept maps are quite different in format but similar in function. Each provides a visual structure that attempts to assure that all relevant information and relationships are considered, that focuses attention on pivotal concerns, and that identifies strengths and weaknesses. Popular memory aids such as underlining, note-taking, and paraphrasing do not fulfill these objectives as reliably.

The scientist who attempts to visualize the entire pattern of evidence risks neglecting a crucial relationship. Writing it in a systematized form may reveal that gap.

Because model/observation tables, outlines, and concept maps compel the scientist to organize knowledge, they are a wonderful first step toward writing up scientific results for publication.

* * *

Confirmation and Refutation of Hypotheses

The evaluation aids can organize a set of evidence effectively. The crux of evidence evaluation, however, is scientific judgment concerning the implications of datasets for hypotheses. Evaluation aids, like scientific progress, constantly demand this judgment: do the data confirm or refute the hypothesis?

Confirmation and **verification** are nearly synonymous terms, indicating an increase in confidence that a hypothesis is correct. Unfortunately, the terms confirmation and verification are widely misused as simple true/false discriminators, like prove and disprove. Rarely are experiments so diagnostic as to prove or disprove a hypothesis. More frequently, evidence yields a qualitative confirmation or its converse, refutation.

It is often said (e.g., by Einstein, Popper, and many others) that no quantity of tests confirming a hypothesis is sufficient to prove that hypothesis, but only one test that refutes the hypothesis is sufficient to reject that hypothesis. This asymmetry is implicit to deductive logic. Two philosophical schools -- justificationism and falsificationism -- begin with this premise and end with very different proposals for how science 'should' treat confirmation and refutation of hypotheses.

* * *

The philosophical school called **justificationism** emphasizes a confirmation approach to hypothesis testing, as advocated by proponents such as Rudolf Carnap. Any successful prediction of an hypothesis constitutes a confirmation -- perhaps weak or perhaps strong. Each confirmation builds confidence. We should, of course, seek enough observations to escape the fallacy of hasty generalization.

Carnap [1966] recommended increasing the efficiency or information value of hypothesis testing, by making each experiment as different as possible from previous hypothesis tests. For example, he said that one can test the hypothesis "all metals are good conductors of electricity" much more effectively by testing many metals under varied conditions than by testing different samples of the same metal under rather similar conditions. This approach is analogous to the statistical technique of using a representative sample rather than a biased one, and its goal is the same: to assure that the properties exhibited by the sample are a reliable guide to behavior of the entire population.

Carnap seems to take this analogy seriously, for he argued that it is theoretically possible to express confirmation quantitatively, by applying a 'logical probability' to each of a suite of hypothesis tests and calculating a single 'degree of confirmation' that indicates the probability that a hypothesis is correct. Jeffrey [1985] proposed adoption of 'probabilistic deduction', the quantitative assessment of inductive arguments, based on calculating the odds that a hypothesis is correct both before and after considering a dataset.

Justificationism and probabilistic deduction have been abandoned by philosophers of science and ignored by scientists, for several reasons. The decision on how many observations are needed is, unfortunately, a subjective one dependent on the situation. The quest for heterogeneous experimental conditions is worthwhile, but it is subjective and theory-dependent. Even if we could confine all of our hypothesis tests to statistical ones with representative samples, we cannot know that the tests are representative of all possibly relevant ones. The confirming observations are fallible and theory-dependent; we look mainly for what the hypothesis tells us is relevant. Furthermore, we have no way of knowing whether a different hypothesis might be proposed that explains all of the results just as well. Thus we can infer from a large number of confirmations that a hypothesis is *probably* correct. We cannot, however, quantify this probability or even know that it is greater than 50%.

* * *

Karl Popper focused on these weaknesses of confirmation and concluded that additional 'confirmations' do not necessarily and substantially increase confidence in a hypothesis. In reaction, he created a philosophy for hypothesis testing known as **falsificationism**. Starting from the premise that the only compelling experiment is one that disproves a hypothesis, he argued that the task of science should be falsification, the rejection of false theories.

First proposed in 1920 and eloquently advocated by Popper, falsificationism had a substantial following among philosophers of science for several decades, and many aspects of it survive. Yet falsifiability has been virtually ignored by scientists. Popper's vision of science is generation of a myriad of ideas followed by ruthless falsification and rejection of the majority. This vision does not correspond with the experience of scientists, but of course our subjective experience could be misleading.

Most scientists do agree that testability is a fundamental criterion for deciding which hypotheses are worthy of attention, but none agree with Popper's assessment that falsifiability is supreme, nor that minor supporting roles are played by confirmation, discovery, insight, and subjective context-dependent evaluation. "An idea may be neither demonstrably true nor false, and yet be useful, interesting, and good exercise" [Trotter, 1941]. A concept may be embraced even without falsifiability, if it is capable of finding elegance of pattern among anomalous observations. Virtually the only mention of falsifiability that I have seen in my field (geology/geophysics) was Ken Hsü's claim that Darwinian evolution is nonscientific because it is not falsifiable. Is the scientific method nonscientific, because its assumption of causality is neither provable nor disprovable?

Falsifiability is a tool, not a rule. The logical flaw in falsificationism is its deductive conclusion that a single inconsistent observation disproves a hypothesis. Scientists do not agree to follow this simple path for evaluating hypotheses, because the source of the inconsistency may be problems in the data, assumptions, or experimental conditions. Kuhn [1970], several other philosophers of science, and Wilson [1952] have cited numerous examples of theories surviving in spite of 'falsifying observations':

> Newton's laws exhibited incredible predictive value. Although they failed to account completely for planetary orbits, they were not rejected.
>
> The chemical 'law' of Dulong and Petit is that the specific heat of each solid element multiplied by its atomic weight is approximately 2 calories per degree. This empirical relationship was used for many years, in spite of the early recognition that it did not work for either silicon or carbon. The exceptions were neither ignored nor used to reject the hypothesis. Ultimately they helped to guide establishment of a law more founded in theory. From the perspective of that new law, Dulong and Petit's law was a special limiting case.
>
> Copernicus' 1543 proposal that the earth revolves around the sun initially conflicted with many observations. The 'tower argument' was particularly damning: if the earth really is spinning, then an object dropped from a tower should land west of the tower, not -- as observed -- at its foot. Fortunately the theory was not discarded.

* * *

Power of Evidence

The successful middle ground between avid justificationism and falsificationism is a concern with the power of evidence. *Information is proportional to astonishment,* or, in terms of informa-

tion theory, the value of a piece of information is proportional to the improbability of that information.

The most powerful and therefore most useful experiment depends on the situation: it may be the experiment most likely to confirm, or to refute, a hypothesis. The fate of most novel, sweeping hypotheses is a quick death, so their refutation has little impact on science. Confirmation of such a hypothesis, on the other hand, does have substantial information value. Similarly, many hypotheses are only incremental modifications of previous theories and so their confirmations are expected to be *pro forma*. Refutation of such a hypothesis may force us to rethink and revise our core assumptions. Well established theories are not normally tested directly, but when such a theory is found to be irreconcilable with an apparently rigorous experiment, this powerful and informative anomaly fosters intensive analysis and experimentation.

Ronald Giere [e.g., 1983] is one of the leading proponents of the 'testing paradigm', more popularly known as the **diagnostic experiment**. A diagnostic experiment avoids the ambiguity of weak true/false tests such as those of justificationism and falsificationism, and it avoids qualitative value judgments. The diagnostic test is the key test, the scalpel that cuts to the heart of a hypothesis and yields a result of 'true' if the prediction is confirmed, and 'false' if the prediction is refuted.

For normal science, *the diagnostic experiment is generally a myth -- an ideal to be sought but seldom achieved*. The diagnostic experiment is, nevertheless, a worthy goal, for it is far better to fall short of the perfectly diagnostic experiment than to fire random volleys of experiments in the general direction of an hypothesis.

> Jonas Salk [1990] has the ideal of a diagnostic experiment in mind when he says: "Solutions come through evolution. It comes from asking the right question. The solution preexists. It is the question that we have to discover."

> Clausewitz [1830] gives analogous advice to military planners: "A certain center of gravity, a center of power and movement, will form itself, on which everything depends. . . We may, therefore, establish it as a principle, that if we can conquer all our enemies by conquering one of them, the defeat of that one must be the aim of the War, because in that one we hit the common center of gravity of the whole War."

* * *

The **Raven's Paradox** [e.g., Lambert and Brittan, 1970; Mannoia, 1980] is an inductive problem that provides a surprising and useful perspective on the power of evidence. Suppose we wish to test this hypothesis: 'All ravens are black.' Symbolically, we can express this hypothesis as R⇒B (**R**aven implies **B**lack) or 'R, ∴B' (Raven, therefore Black). Any example of a raven that is black provides confirmatory evidence for the validity of the hypothesis. Even one instance of a raven that is not black proves that the hypothesis is wrong.

The paradox arises when we consider the implications of the following rule of logic: each statement has logically equivalent statements (Chapter 4), and if a statement is true, its logically equivalent statement must also be true. A logical equivalent of the hypothesis 'All ravens are black' is 'All non-black things are not ravens.' Caution (or practice) is needed to be certain that one is correctly stating the logical equivalent. 'All non-ravens are not black' superficially sounds equivalent to 'All ravens are black,' but it is not.

The Raven's Paradox is this: anything that is both not black and not a raven helps confirm the statement that all ravens are black. Without ever seeing a raven, we can gather massive amounts of evidence that all ravens are black.

The Raven's Paradox has been the subject of much discussion among philosophers of science. Some of this discussion has concluded that seemingly absurd types of evidence (not-Black + not-Raven confirms R⇒B) are nevertheless valid, but most arguments have centered on the intrinsic weakness of the confirmation process. In contrast, I see the tests of the Raven's Paradox, like all scientific evidence, in terms of information value. Observations of non-ravens do help confirm the hypothesis that 'All ravens are black,' but the information value or evidential power of each observation of a non-raven is miniscule. Even thousands of such observations are less useful than a single observation of a raven's color. Were this not so, we could use the concept of logical equivalence to 'confirm' more outrageous hypotheses such as 'All dragons are fierce.'

Like the example in Chapter 3 of the 'cause' of Archimedes' death, many inferences form a pattern: $X_1 \Rightarrow X_2 \Rightarrow X_3 \Rightarrow X_4$. All elements of the pattern are essential; all elements are not of equal interest. Familiar relationships warrant only peripheral mention. The pattern link of greatest scientific interest is the link that has the maximum information value: the most unusual segment of the pattern.

* * *

Scientific research is intimately concerned with the power of evidence. *Inefficient scientists are transient scientists*. The demand for efficiency requires that each researcher seek out the most powerful types of evidence, not the most readily available data. In the case of the Raven's Paradox, this emphasis on experimental power means first that only ravens will be examined. Furthermore, a single instance of a non-black raven is much more important than many instances of black ravens, so the efficient scientist might design an experiment to optimize the chance of finding a non-black raven. For example, the hypothesis 'All dogs have hair' could be tested by visiting several nearby kennels, but a single visit to a Mexican kennel, after some background research, might reveal several examples of Mexican hairless dogs.

To the logician, a single non-black raven disproves 'All ravens are black', and a single Mexican hairless disproves 'All dogs have hair.' The scientist accepts this deductive conclusion but also considers the total amount of information value. If exceptions to the hypothesis are rare, then the scientist may still consider the hypothesis to be useful and may modify it: '99.9% of ravens are black and 0.1% have non-black stains on some feathers,' and 'All dogs except Mexican hairlesses have hair.'

* * *

Hypothesis Modification

The distinction between scientists' and logicians' approaches does not, of course, mean that the scientist is illogical. Confirmation and refutation of hypotheses are essential to both groups. They usually do not, however, lead simply to approval or discarding of scientific hypotheses. In part, this outcome is progressive: the hypothesis as originally stated may be discarded, but the scientific companion of refutation is modification. In many cases, simple acceptance or rejection is not possible, because hypotheses are usually imperfect.

Confirmation or falsification of a hypothesis, like the 'diagnostic experiment', can be difficult to achieve, for several reasons:

- Many hypotheses have inherent ambiguities that prevent simple confirmation or falsification. An experiment may favor one interpretation of a hypothesis, but the door is left open for other interpretations.

- Most experiments, in spite of careful experimental design, have at least some inherent ambiguity.
- Most hypotheses and their tests have associated assumptions and concepts. Refuting evidence indicates inadequacy of either the main hypothesis or corollaries, and one may not know confidently which to reject. Typically, the 'hard core' of a theory is relatively invulnerable to attack, and we refute or modify the 'protective belt' of ancillary hypotheses, assumptions, and conditions [Lakatos, 1970].
- Instead of directly testing a hypothesis, we usually test deductive or inductive predictions derived from the hypothesis. This prediction may be wrong rather than the hypothesis.

Proof or disproof of a hypothesis is often impossible; rarely, search for proof or disproof can be undesirable. Frequently the scientific community loses interest in endless tests of a hypothesis that is already judged to be quite successful; they change the focus to characterization of the phenomenon. Then inductive predictions are the target of experiments, because little ambiguity remains about whether one is testing the hypothesis or its inferred implications. Symbolically, if h is a hypothesis, p_i is an inductive prediction, and p_d is a deductive prediction, then some possible hypothesis tests are:

h, directly testable;

$h \Rightarrow p_d$, p_d testable so h testable;

$h \Rightarrow p_i$, p_i testable but h is not directly tested.

A school of thought known as conventionalism recognizes the networked nature of most hypotheses and the associated ambiguity of most confirmation/refutation evidence, as well as the seductiveness of modifying an otherwise successful hypothesis to account for inconsistent observations. Conventionalists conclude that subjective judgment is required in evaluating hypotheses, and they suggest that values such as simplicity and scope are used in making these judgments.

If the conventionalists are correct about how science works, then the subjectivity of evidence evaluation is a major obstacle to our quest for reliable knowledge. The weakness of conventionalism is its fluidity. Two scientists can examine the same evidence and embrace opposing views, because of different criteria for evidence evaluation. Most hypotheses are wrong, but demonstration of their errors leads more often to a modification of the hypothesis than to its rejection. This band-aid approach, though powerful and often successful, can lead the researcher into evaluating how reasonable each slight modification is, without detecting how cumbersome and unreasonable the composite hypothesis has become. Unless one holds tightly to the criterion of simplicity, there is the danger that any wrong hypothesis will stay alive by cancerously becoming more and more bizarre and convoluted to account for each successive bit of inconsistent data.

> When Galileo aimed his telescope at the moon and described mountains and craters, his observations conflicted with Aristotelian cosmology, which claimed that all celestial objects are perfect spheres. A defender of the old view had this *ad hoc* explanation: an invisible, undetectable substance fills the craters and extends to the top of the mountains.

Imre Lakatos [1970] attempted to put a brake on this unconstrained *ad hoc* hypothesis modification by imposing a standard: if a hypothesis is modified to account for a conflicting observation, then it must not only account for all previous results just as well as did the original hypothesis, but also make at least one new and successful prediction.

Lakatos' goal is worthwhile: steering the evolution of hypotheses toward those that have greater explanatory power. His method is feasible, if a bit awkward. Usually the hypothesis revision occurs after a project has obtained its research results, so the actual test of the new prediction is deferred for a later paper by the same or different authors. Lakatos' criterion is virtually unknown and unused among scientists, however. Its problem is the same as that of falsificationism: it is an outside judgment of what scientists *should* do (according to the proponent), rather than a description of what they *actually* do, and we scientists are not persuaded of the need to change. We can, however, be alert for *ad hoc* hypotheses, and we do expect a modified hypothesis to explain more than its predecessor.

I and many other scientists are close to this conventionalist view. We are, perhaps, even closer to Thomas Kuhn's perspective, described in the next section.

* * *

Paradigm and Scientific Revolution

Thomas Kuhn's 1963 (and 1970) book The Structure of Scientific Revolutions overthrew our perception of scientific change. We had imagined scientific change as a gradual process, involving incremental advancement in techniques, evidence, and hypotheses, which resulted in a steady increase in scientific knowledge.

Our textbooks reinforced this view by portraying the history of scientific thought from our present perspective. Early ideas are judged to be important and relevant only to the extent that they contribute to the continuous evolution toward the current ideas. Textbooks express the outcomes of scientific revolutions as discoveries of new ideas; they avoid confusing this picture with discussion of the process of scientific upheavals and of the ideas that have been superseded. Because most science students read textbooks rather than scientific articles prior to initiating their own graduate research, their perception of scientific change is fossilized even before they have a chance to contribute to that change.

Kuhn said that we must consider scientific results in the context of the sociological factors and scientific perspectives of their time. He saw the advance of science more as a staircase than a ramp. Within each scientific field, long periods of stability and consolidation are followed by short periods of major conceptual revision, or *paradigm change*. I think that this view of science is progressive: not only is it a more realistic perspective, but also it offers insights into which scientific methods are most appropriate at different points in the evolution of a science.

A **paradigm** is a suite of "universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners" [Kuhn, 1970]. Kuhn realized that this definition is vague and sloppy. To me, a paradigm is a coherent suite of theories or concepts that guide interpretations, choice of relevant experiments, and development of additional theories in a field or discipline. Physics paradigms, for example, included Newtonian dynamics, general relativity, and quantum mechanics.

We can understand paradigms better by considering a field in its pre-paradigm state. Data collection is unfocused, a fishing expedition rather than a hunter's selection of prey. Facts are plentiful, but the overall patterns and organizing principles are unclear. Several schools of thought compete, none agreeing on what phenomena warrant study and none providing broad-scope hypotheses. Research is overwhelmed by the apparent complexity of the subject.

* * *

When a paradigm guides a scientific field, nearly all research is considered in relation to that paradigm. Research is focused; the paradigm indicates which research topics are appropriate and worthwhile. Both theoretical and experimental studies are largely confined to three foci:

1) collecting data to test predictions of the paradigm;

2) pursuing aspects that may elucidate seminal phenomena. These investigations often require development of more sophisticated, more accurate equipment; and

3) attempts to 'articulate' the paradigm, including efforts to extend it and account for other phenomena, and attempts to resolve apparent problems or ambiguities.

Paradigm change is rare; working under a guiding paradigm is the norm. These 'mopping-up operations' are exciting because they promise goal-oriented, steady progress rather than a frustrating floundering. Often the results of experiments are readily predictable, but the work is still challenging. Ingenuity and insight are needed to determine how to conduct the experiment most successfully and elegantly.

* * *

Researchers ignore most data that appear to be unrelated to or unexplained by the paradigm. Moreover, we tend to ignore evidence that conflicts with the paradigm. No paradigm explains all observations, because no paradigm provides ultimate and final truth. Yet the immense explanatory power of the paradigm leads scientists to think of the contradictory data either as mistaken or as explicable by future elaborations of the paradigm. In either case, the results can be ignored for the moment -- or so we tell ourselves, if we even notice the contradictions. Publication of evidence that seems to conflict with the paradigm is hazardous, for the authors risk being branded as nonbelievers or outsiders.

An established paradigm is insulated from overthrow, by both the tendency to ignore discrepant facts and by the habit of refining hypotheses and paradigms [Kuhn, 1970]. Even when many anomalies are found, we do not discard the paradigm, for rejection leaves a vacuum. Rejection implies that all the predictive successes of the paradigm were coincidental. Only when a new potential paradigm appears will abandonment of the old be considered. Scientific inertia is conservative: a new paradigm is accepted only if it is demonstrably superior -- not merely equal in success -- to the old paradigm.

Timing of the new paradigm's appearance is critical. It must be considered when anxiety over anomalies in the old paradigm is high. Without the leverage of anomaly anxiety, attempts to challenge the paradigm's authority are likely to fail (e.g., Plato vs. democracy, Aristotle vs. slavery, Descartes vs. experimental science, and Einstein vs. quantum mechanics). Introduction of a new theory too early will encounter complacency with the old one. Indeed, sometimes the new paradigm is a reintroduction and slight refinement of a previously proposed idea, which had failed to gain momentum.

> Discovery "commences with the awareness of anomaly (i.e., with the recognition that nature has somehow violated the paradigm-induced expectations), continues with extended exploration of the area of anomaly, [and] concludes when the paradigm has been adjusted so that the anomalous has become the expected." [Kuhn, 1970]

* * *

Paradigm change begins with a single anomaly that cannot be ignored. Anomaly creates a sense of trauma or crisis, as we saw in the card-flashing experiment [Bruner and Postman, 1949] when

the subject said, "I'm not even sure now what a spade looks like. My God!" The sense of crisis and anxiety grows with recognition of the many smaller anomalies that had been overlooked. The entire foundation of the field seems unstable, and doubts arise about the value of familiar paradigm-inspired experiments.

Anxiety creates a willingness, even a need, to consider alternative paradigms. The field splits into two camps: that which suggests competing piecemeal solutions to the various anomalies, clinging to the old paradigm, and that which considers alternative paradigms. This second group of investigators refuses to accept rationalizations of the major anomaly. These scientists explore the anomaly more deeply and attempt to characterize it, simultaneously looking for invalid assumptions.

> "That is in the end the only kind of courage that is required of us: the courage to face the strangest, most unusual, most inexplicable experiences that can meet us." [Rilke, 1875-1926]

Reconciliation of the problems seldom comes rapidly. The period of paradigm crisis can last for years or decades, and anxiety may become discouragement. Perhaps the new paradigm will require new technology and its attendant new insights. Almost always, the new paradigm is discovered by someone young or new to the field, someone less hampered than most by perspectives and assumptions of the old paradigm. The new paradigm may be a radical modification of the old paradigm. The old paradigm may be seen as a special limiting case of the new one, as was the case for Newtonian dynamics when seen from the perspective of Einstein's dynamics.

Paradigm change may be led by a few people, but usually it involves many people working over a period of several years. Within the subgroup that had been bothered by the anomalies, a consensus of both experimenters and theoreticians emerges, concerning the advantages of a new paradigm over the old one. Simultaneous independent discoveries are likely. Now is the most exciting time, with the (mostly young) proponents of the new paradigm exploring the range of its applications. The pace of change is extremely fast: only those who are attending conferences, receiving preprints, and learning the new jargon are fully aware of these changes.

Polarization of old and new schools continues well beyond the acceptance by the majority of the new paradigm. The old and new paradigms identify different subjects as appropriate for research and emphasize controlling different variables. Communication between the two schools breaks down. Neither paradigm accounts for every observation; thus each group can point to anomalies or weaknesses in the other paradigm. But with time the demand of the new majority is fulfilled: "convert or be ignored" [Kuhn, 1970].

* * *

These interpretations of the pattern of change in science are those of Thomas Kuhn; they are not accepted universally. Stephen Toulmin [1967] suggested that scientific change is more evolutionary than Kuhn has pictured it. Toulmin used the analogy of biological evolution, emphasizing that competing theories abound, and the more successful ones eventually triumph. The analogy was unfortunate, for most paleontologists now see evolution as dominantly episodic or revolutionary -- a punctuated equilibrium [Eldredge and Gould, 1972].

Scientific change is punctuated illumination. For both scientific and biological evolution, relatively stable periods alternate with periods of dynamic change, but no one suggests that the stabler times are stagnant. Mannoia [1980] summarized the philosophical trend of the seventies as moving away from the Kuhn/Toulmin perspectives, toward an 'historical realism', but I think that the jury is still out.

Few books in philosophy of science have attracted the interest of scientists; Kuhn's [1970] book is an exception. His vision of scientific change -- continuous incremental science, plus rare revolution -- is fascinating to those in fields undergoing this punctuated illumination. The changes associated with the 1968 geological paradigm of plate tectonics appear to fit his model, as does the recent paradigm of chaos, described by Gleick [1987] as a "revolution in physical sciences". Successful prediction always confirms a model more persuasively than does detection of apparent pattern within existing data.

* * *

> "Yet when we see how shaky were the ostensible foundations on which Einstein built his theory [of general relativity], we can only marvel at the intuition that guided him to his masterpiece. Such intuition is the essence of genius. Were not the foundations of Newton's theory also shaky? And does this lessen his achievement? And did not Maxwell build on a wild mechanical model that he himself found unbelievable? By a sort of divination genius knows from the start in a nebulous way the goal toward which it must strive. In the painful journey through uncharted country it bolsters its confidence by plausible arguments that serve a Freudian rather than a logical purpose. These arguments do not have to be sound so long as they serve the irrational, clairvoyant, subconscious drive that is really in command. Indeed, we should not expect them to be sound in the sterile logical sense, since a man creating a scientific revolution has to build on the very ideas that he is in the process of replacing." [Hoffmann, 1972]

* * *

Pitfalls of Evidence Evaluation

Scientific progress under a guiding paradigm is exhilarating. Paradigm-driven science can, however, undermine the objectivity with which we evaluate hypotheses and evidence.

Hidden Influence of Prior Theory on Evidence Evaluation

Data evaluation should consist of three separate steps: (1) objective appraisal of the observations, (2) confirmation or refutation of a hypothesis by these data, and (3) overall evaluation of a hypothesis in the context of these and other observations. All too often, we allow our prior opinion of a hypothesis to influence the evaluation of new evidence (steps #1 & 2), without being aware of the bias. This hidden influence is a pitfall, whereas it is completely valid to weight prior evidence more than the new data (step #3). In both cases, the impact of evidence depends on the perceived strength of the hypothesis it affects. Evidence sufficient to uproot a weakly established hypothesis may fail to dislodge a well established one.

We value simplicity, and it is much simpler and more comfortable if new evidence confirms previous beliefs than if it creates conflict. Ideally, one (and only one) hypothesis is consistent with all observations. To obtain this ideal, we may subconsciously reject evidence that conflicts with the hypothesis, while overemphasizing evidence that supports it. We must beware this subconscious theory-based rejection of data.

Children and adults use similar strategies to cope with evidence that is inconsistent with their prior beliefs [Kuhn et al., 1988]:

- consciously recognize the discrepancy and conclude that either the hypothesis or the evidence is wrong;
- consciously recognize the discrepancy, then deliberately revise the hypothesis to make it more compatible with the evidence;
- reduce the inconsistency by biased interpretation of the evidence;
- subconsciously revise the hypothesis to make it more compatible with the evidence.

All four strategies also are employed by scientists, but only the first two are valid. The first three have been discussed already and are also familiar in daily experience. Subconscious revision of a hypothesis, in contrast, is a surprising pitfall. Kuhn et al. [1988] found that subjects usually modified the hypothesis *before* consciously recognizing the relationship of the evidence to the hypothesis. They seldom realized that they were changing the hypothesis, so they failed to notice when their theory modification was implausible and created more problems than it solved. Fortunately for science but unfortunately for the scientist who succumbs to the pitfall of subconscious hypothesis modification, someone usually detects the error.

Kuhn et al. [1988] found that hypotheses of causal relationships between variables are particularly resistant to overthrow by new data. The new data must overcome the expectation of a correlation; even if the data set as a whole does so, nonrepresentative subsets may still appear to confirm the correlation. Furthermore, the original proposal of a causal relationship probably also included a plausible explanation. To discard the correlation is also to reject this explanation, but the new data do not even address that argument directly.

The hidden influence of accepted hypotheses on evidence evaluation harms scientists as well as science. A scientist's beliefs may fossilize, leading to gradual decrease in creative output (though not in productivity) throughout a professional career.

As we saw in the previous section on paradigms, hidden influence of prior theory has *other manifestations: (1) ignoring data inconsistent with the dominant paradigm; (2) persistence of theories in spite of disproof by data; and (3) failure to test long-held theories*.

* * *

Incremental Hypotheses and Discoveries

Because the dominant paradigm molds one's concepts, it largely controls one's expectations. Hypotheses and discoveries, therefore, tend to be incremental changes and elaborations of the existing theories, rather than revolutionary new perspectives. Mannoia [1980] says that "the answers one obtains are shaped by the questions one asks."

* * *

'Fight or Flight' Reaction to New Ideas

The expression 'fight or flight' describes the instinctive reaction of many animal species to anything new and therefore potentially threatening. Beveridge [1955] pointed out that 'fight or flight' is also a scientific pitfall. When presented with new ideas, some individuals fight: the theory is immediately rejected, and they only listen to pick out flaws. Their biased attitude should not be confused with the scientifically healthy demand, 'show me', suspending judgment until the evidence is heard. Other scientists flee, ignoring any new idea until more conclusive, confirming evidence can be provided. A scientist who rejects relevant evidence, on the grounds that it leaves ques-

tions unanswered or it fails to deliver a complete explanation, is confusing the responsibilities of evidence and hypothesis.

> "The mind likes a strange idea as little as the body likes a strange protein, and resists it with a similar energy. . . If we watch ourselves honestly we shall often find that we have begun to argue against a new idea even before it has been completely stated." [Trotter, 1941]

> "In the 1790's, philosophers and scientists were aware of many allegations of stones falling from the sky, but the most eminent scientists were skeptical. The first great advance came in 1794, when a German lawyer and physicist, E.F.F. Chladni, published a study of some alleged meteorites. . . Chladni's ideas were widely rejected, not because they were ill conceived, for he had been able to collect good evidence, but because his contemporaries simply were loathe to accept the idea that extraterrestrial stones could fall from the sky." [Hartmann, 1983]

* * *

Confusing the Package and Product

Scientists are not immune to the quality of the sales pitch for a set of evidence. Unless the reader is put off by blatant hype, the sales pitch exerts a subconscious influence on one's evaluation of the evidence. For example, consider the statement "All hypotheses are wrong, but some are more wrong than others." Catchy expressions tend to go through one's head and thereby gain strength, while qualifications and supporting information are forgotten. In this one a defeatist mood is enforced, rather than the optimistic prospect of growth and evolution of changing ideas. To separate the objective evidence from the effects of presentation style, paraphrasing arguments can help.

* * *

Pitfall Examples

> For over 2000 years, from the ancient Egyptian, Greek, Roman, Chinese, and Japanese cultures to the 19th century, there persisted the myth of the oxen-born bees. The myth, 'confirmed' by observation, explained that decaying carcasses of oxen transformed into a swarm of honeybees.
>
> The birth that people witnessed so often was not of honeybees but rather of the fly Eristalis tenax, which looks similar. The flies do not generate spontaneously; they hatch from eggs laid in the carcasses. In all that time, though people had seen developing honeybees in honeycombs, no one captured the oxen-born bees and attempted to raise them for honey, nor did they compare them with honeybees, nor did they observe the egg laying or the eggs [Teale, 1959].

Pitfalls: failure to test long-held theories;

missing the unexpected;

missing important 'background' characteristics.

> In 1887, physicist Albert Michelson and chemist E.W. Morley carried out an experiment to detect the earth's motion through the ether. They measured the difference in the travel times of light moving at different angles to the earth's presumed direction through the ether. Although theory indicated that the measurements were sensitive enough to detect this effect, the Michelson-Morley experiment found no difference. Fortunately for physics, these scientists did not suppress their negative results. They published, although for 15 years Michelson considered the experiment a failure

[Hoffmann, 1972]. Theories assuming the existence of an ether survived the emergence of this and other anomalies, until Einstein's 1905 paper on special relativity changed the paradigm and accounted for the Michelson-Morley results.

Pitfalls: theories persist even when disproved by data;

ignoring data inconsistent with dominant paradigm.

In the section called 'Paradigm and Scientific Revolution' in this chapter, Jarrard gives a detailed interpretation of Thomas Kuhn's ideas, yet he dismisses Stephen Toulmin's arguments by attacking his analogy, and he dismisses alternative opinions with a single reference.

Pitfalls:

ignoring data inconsistent with dominant paradigm;

advocacy masquerading as objectivity;

biased evaluation of subjective data.

Chapter 8: Insight

[Watterson, 1993]

"To see a World in a Grain of Sand
And a heaven in a Wild Flower
Hold Infinity in the palm of your hand
And Eternity in an hour."
[Blake, ~1803]

Rain fell in the mountains, and a brook was born. The young brook splashed and slithered over rocks and under branches, meeting other brooks and merging with them, growing and slowing. No longer a brook, a powerful river emerged from the mountains in a final waterfall, which encountered the desert.

"You cannot pass," said the desert. But nothing had ever stopped the river, so it flowed forward, now out across the desert. The desert passively soaked up the river's water, stopping the river's advance. "You cannot pass," said the desert.

The river had learned persistence. It continued to flow out into the desert, expecting eventually to win passage. But the desert was no stranger to persistence.

The river accumulated floating debris at the mouth of the waterfall, forming a temporary dam, building up a huge backlog of water, then bursting upon the desert. The desert seemed to be overwhelmed by the torrent, but only temporarily.

The river tried to avoid the desert, skirting it by flowing along the base of the mountain's foothills. The desert found and drank the river water.

The river felt defeated. Persistence, power and avoidance had always succeeded before but had failed to overcome this obstacle. Everything the river had ever been was for naught - its youthful brooks, later streams, and final river strength - in the face of this obstacle. Everything? Or was there a time before the brooks?

The river gave itself to the wind. And the water that fell over that final waterfall never touched the desert floor. The water was swept up and evaporated, carried by the wind out of reach of the desert, across the desert, to mountains beyond.

Rain fell in the mountains, and a brook was 'born'.

[loosely based on a Sufi teaching story, e.g., Shah, 1970]

* * *

Role of Insight in Science

Hypothesis and observation, or theory and empiricism, are only two of the three essential ingredients of modern scientific method. The third pillar of wisdom is insight -- the sudden transcendence of obstacles by a new perspective, like the river transforming to vapor and crossing the desert. Insight changes the counterpoint of hypothesis and data into the upward spiral of hypothesis, data, insight, new hypothesis, different data, . . .

Other terms are used synonymously with 'insight': illumination, intuition, serendipity, scientific hunch, revelation, inspiration, enlightenment, sudden comprehension, guess, and discovery. Most of these terms have such a heavy connotation of either religious, psychological, or everyday secular meaning, however, that they are somewhat distracting to use in the current scientific discussion.

Insight brings joy to science. Without this thrill, many of us would not be scientists.

In their excellent and still timely article on the role of 'scientific hunch' in research, Platt and Baker [1931] define a scientific hunch as "a unifying or clarifying idea which springs into consciousness suddenly as a solution to a problem in which we are intensely interested. . . A hunch springs from a wide knowledge of facts but is essentially a leap of the imagination, in that it goes beyond a mere necessary conclusion that any reasonable man must draw from the data at hand. It is a process of creative thought." This is not deduction, but induction -- and sometimes induction totally unwarranted from the available evidence. Sometimes it is a solution to a minor technical problem, and sometimes an insight so fundamental that we can never again see the world in the old way.

Helmholtz [1903], Wallas [1926], Platt and Baker [1931], Sindermann [1987] and others concisely describe scientific method as consisting of **four stages: preparation, incubation, illumination, and verification**. I agree that these four stages are real, and we will come back to them soon in considering how insight can be encouraged or hampered. These four terms betray a strong bias, however, toward casting illumination as the central and most important aspect of science, with the other stages serving only a supporting role. Such a view is by no means universal. Indeed, when 232 scientists replied to a questionnaire concerning insight in science, 17% said that scientific revelations or hunches never help them find a solution to their problems [Platt and Baker, 1931]. I suspect that they rely on insight as much as I do, but they dislike the connotations of the words 'revelation' and 'hunch', and they prefer to think of science as more rational than those terms imply. Possibly also, they shy from insight's 'nonscientific' characteristics: it is nonreproducible, nonquantifiable, unpredictable, unreliable, and sometimes almost mystical.

Scientists' reliance on insight is incredibly diverse, partly because of variations in ability but also largely because of value judgments concerning rational data-gathering versus irrational insight. Some get ideas and experiment to test their ideas, some prefer to test others' hypotheses, and some try to gather data until an answer emerges as virtually proved. A few types of research claim to thrive on minimal insight. For example, C.F. Chandler said that one could solve any problem in chemical research by following two simple rules: "To vary one thing at a time, and to make a note of all you do" [Platt and Baker, 1931]. To many scientists, such an approach is either infeasible or boring.

The four stages of research occupy unequal proportions of our research time. We might wish that insight were 25% of the job volumetrically as well as conceptually, but Thomas Edison's generalization is probably more accurate: "science is 99% perspiration and 1% inspiration." At least that is what a former advisor told me when, as a new graduate student, I showed little enthusiasm for spending countless hours doing routine measurements for his project.

"Before illumination, carry water, chop wood.
After illumination, carry water, chop wood." [Zen saying]

The content of the *preparation* and *verification* stages is primarily routine, straightforward, and mechanical, requiring different skills than are needed for the insight stage. Courses and books (including this one) usually devote much more attention to these skills than to techniques for enhancing insight.

The advertising industry often uses a technique known as anxiety/relief: first create an anxiety, and then offer your product as a potential relief. Today part of the success of this technique is attributable to its strongly conditioned pattern. Why do people like myself get addicted to high anxiety jobs? Perhaps it is because the solutions, when found, are that much sweeter. Problem solving may fulfill a similar role in science, as a non-threatening pattern of anxiety and relief. And the intensity of the thrill of insight may depend partly on the duration and intensity of the quest that preceded it.

* * *

Characteristics of Insight

Insight occupies a continuum from conscious to unconscious, from minor problem-solving to mystical experience. Always it involves a leap beyond the available evidence, to unforeseen paths. Almost always it brings a sense of certainty, a dangerous conviction of the truth of the insight.

Poincaré [1914] describes insight's "characteristics of conciseness, suddenness and immediate certainty." Another typical characteristic is joy or exhilaration. We will return to the characteristic of immediate certainty in a later section on insight pitfalls. The following descriptions of insights illustrate both their variety and some of their common elements:

> "He who has once in his life experienced this joy of scientific creation will never forget it; he will be longing to renew it." [Kropotkin, 1899]
>
> "The joy of discovery is certainly the liveliest that the mind of man can ever feel." [Bernard, 1865]
>
> "It came to me in a dream and it's money in the bank. It's so simple it's ridiculous. . . Read it and weep." [1990 fax from a colleague who is an electronics technician, describing a new equipment design]
>
> Alfred Russel Wallace [1853], who discovered evolution independently of Charles Darwin, described his walks in the Welsh countryside: "At such times I experienced the joy which every discovery of a new form of life gives to the lover of nature, almost equal to those raptures which I afterwards felt at every capture of new butterflies on the Amazon."
>
> Albert Einstein, in a 1916 letter [cited by Hoffmann, 1972], described his discovery and confirmation of general relativity after an 11-year search: "Imagine my joy at the feasibility of the general covariance and at the result that the equations yield the correct perihelion motion of Mercury. I was beside myself with ecstasy for days."

Of course, insight is neither limited to science nor always best described by scientists:

Author Thomas Wolfe [1936] described his creation of three books as follows: "It was a progress that began in a whirling vortex and a creative chaos and that proceeded slowly at the expense of infinite confusion, toil, and error toward clarification and the articulation of an ordered and formal structure. . . With a kind of hurricane violence that could not be held in check, . . . the storm did break . . . It came in torrents, and it is not over yet."

"The flame of conception seems to flare and go out, leaving a man shaken, and at once happy and afraid. There's plenty of precedent of course. Everyone knows about Newton's [apocryphal] apple. Charles Darwin said his Origin of Species flashed complete in one second, and he spent the rest of his life backing it up; and the theory of relativity occurred to Einstein in the time it takes to clap your hands. This is the greatest mystery of the human mind -- the inductive leap. Everything falls into place, irrelevancies relate, dissonance becomes harmony, and nonsense wears a crown of meaning." [writer John Steinbeck, 1954; cited by Calvin, 1986]

* * *

Conditions Favoring Insight

Perhaps the most valuable result of Platt and Baker's [1931] survey of scientists was its recognition that certain conditions favor achievement of insight:

• **Define the problem**. The more specific one can be in identifying the paradox or problem, the better is one's chance of success. Describing the problem to others sometimes helps, because it forces the researcher to define the problem simply. Sometimes one can solve the larger problem piecemeal by obtaining confident solutions for components of the problem. Yet discrepant observations must not be overlooked. Do the partial solutions suggest that other facts are needed, do they suggest analogies, or do they have an impact on other partial solutions or facts? An exam-taking strategy can be useful here: start with the easiest problems, then work up to the harder ones. This strategy helps build momentum and confidence and it avoids overwhelming the researcher with the magnitude of the problems.

• Complete the initial stage of **preparation**. Killeffer [1969] calls this step accumulation, emphasizing the role of accumulating needed facts. One cannot expect to solve the problem unless the relevant information is available and comprehended. Furthermore, the facts must be organized. Indeed, the juxtaposition of certain facts can provide the mental connection needed for insight, so it may be worthwhile to try arranging the facts in different ways. Sketching or outlining the relationships may help. Mental images may help. Many scientists find that writing a scientific paper triggers insights, because it forces us to organize data, assumptions, and inferences much more systematically than we do mentally. Sometimes one of our assumptions is the obstacle to insight; deliberately listing and challenging all assumptions may help. In summary, preparation includes accumulation, comprehension, evaluation, and organization of data, assumptions, and inferences.

• **Desire a solution**. Having a personal stake in a problem can help or hinder insight; usually it is a strong asset. Preoccupation with the quest keeps the problem churning through one's conscious thoughts and subconscious, providing the needed stage of incubation. Desire for a solution becomes counterproductive if it leads to distracting worry and anxiety. Thus some researchers are

more successful in achieving insights concerning other people's problems than in solving their own problems.

> "The unconscious work goes on only over problems that are important to the waking mind, only when the mind's possessor worries about them, only when he cares, passionately." [Gerard, 1946]

• **Relax and temporarily abandon the problem**. Insight can be fostered as easily as this: simply pause for thought whenever you encounter anomalous data in your research or reading. Even more conducive conditions are the combination of mental relaxation with either physical relaxation or mild exercise [Platt and Baker, 1931]: walking on a beach or in the forest or between work and home, taking a bath, relaxing in bed just before falling asleep or just after awakening. Receptivity is needed to achieve the goal. Abel [1930] said:

> "It is an old saying ever since Archimedes [with the cry, 'Eureka!'] solved the problem of specific gravity in his bath tub. . . that discoveries are not made in the laboratories but in the bath tub, or during an afternoon or evening walk as in the case of Helmholtz, or in the watches of the night when the puzzled brain will not quiet down."

> Charles Darwin [1876] described his discovery of evolution by natural selection as follows: "I can remember the very spot in the road, whilst in my carriage, when to my joy the solution occurred to me."

> "Did not one of the great masters attain enlightenment upon hearing the splash of his own turd into the water?" [Matthiessen, 1978] I don't know, but I doubt it.

> "The fact that the attack is seemingly unsuccessful shows that something is *wrong*. Sometimes merely more information is required. Often, however, the difficulty arises from an incorrect interpretation of the facts at hand. . . In taking up any problem after a period of rest, we have the chance of leaving behind an erroneous point of view and of seizing upon one more fruitful." [Platt and Baker, 1931]

> "The archer hitteth the mark partly by pulling, partly by letting go." [ancient Egyptian saying, cited by Leuba, 1925]

Apparently the subconscious is set working on a problem by our conscious thought and desire for a solution. It keeps working on the problem, trying out possible patterns even when (or especially when) the conscious mind has relaxed and stopped feeding it a variety of distracting extraneous facts.

> As spring comes to the Arctic, the icebound rivers appear to be immune to the warming. Invisibly but pervasively, the ice slowly succumbs to spring's warmth. Without warning, in a few deafening seconds all of the river ice breaks up and begins to flow. The pace of insight is like this breakup.

Additional circumstances, related to the four above, also favor insight. For example, Beveridge [1955] emphasizes the value of discussing ideas with other people. They have different perspectives, and one may benefit from those perspectives or from combining one's knowledge with theirs. Their questions, as well as our need to frame answers in the context of their backgrounds, may force us out of the rut of established thought patterns and into a more fruitful perspective. They or we may spot faulty reasoning, during the explanation of things normally take for granted. Perhaps

the discussion with others will not lead directly to a solution, but it will increase enthusiasm or at least decrease discouragement at how intractable the problem seems to be. Discussions are most likely to encourage insight if they are carried out in a relaxed and friendly, rather than highly critical and defensive, atmosphere.

The Incomplete Guide to the Art of Discovery, by Oliver [1991], suggests that the best way to foster insight is to "try to become associated with the fresh new observations. That is where the discoveries are most likely." Oliver emphasizes that almost every really novel kind of observation brings surprises and enhances understanding. "We need only to recognize an important unexplored frontier and then plan and carry out a sound program of observation of that frontier." Simple!

Most insights illuminate merely the central idea, then the mind rapidly grasps all of the details and implications [Platt and Baker, 1931]. At other times, insights can be partial or fleeting. Many scientists find that it is valuable to jot down such ideas for further consideration later; a pad and pencil near the bed can be helpful.

* * *

Obstacles to Insight

Some obstacles to insight are obvious; others are more insidious, masquerading as an essential part of scientific activity:

• **Distractions** -- particularly unpleasant distractions such as domestic or business worries, anxiety, and fatigue -- destroy the receptivity needed for insight. I have seen anxiety over possible layoffs cut worker productivity by about 50% and cut discoveries by nearly 100%, although management expected that 10% layoffs would cause only 10% reduction in overall productivity.

> In the years 1665-1666, plague in England forced the closing of Cambridge University, so Isaac Newton went home to the village of Woolsthorpe. There, in this brief time, he developed the calculus, discovered the relationship of color to light, and laid the foundation for his later elucidation of the laws of gravitation and dynamics. [Hoffmann, 1972]

> Albert Einstein [1879-1955], whose physics eventually superseded Newton's dynamics, said that the ideal job for a theoretical physicist is to be a lighthouse keeper. In 1933, living in the relatively isolated village of Cromer in England, he said, "I have wonderful peace here; only now do I realize how driven I usually am." On another occasion he expressed similar thoughts about the same location: "I really enjoy the quiet and solitude here. One can think much more clearly, and one feels incomparably better." Yet his most productive period for insights was 1905, during which he worked full-time at the Patent Office.

Pleasant distractions, such as excitement or preoccupation with something other than the immediate problem, can be more insidious but just as inimical to research success and insight. Minor problems, experimental techniques, and equipment modifications are visible and readily attacked forms of problem solving, but also distractions from the main research thrust. Particularly dangerous is the black hole of computers: web crawling, software collection, and software usage can begin as a fascinating and justifiable diversion, then become a time-sink that eclipses their contribution to the primary research objective.

• **Interruptions** are probably the most disruptive type of distraction. Even the expectation of possible interruption is counterproductive to insight, because it prevents total immersion in the problem. Perhaps the most potent step that one can take toward enhancing both productivity and insights is to allot an interruption-free portion of each day or week to thinking about, rather than 'doing', science. Platt and Baker [1931] received many comments such as these on the problem of interruptions:

> "As an example of the benefit due to freedom from interruptions try going to the laboratory on a holiday. Note how easily many formerly complicated problems straighten themselves out, how smoothly the mind works, and how much work is accomplished with little effort."

> "Any employer of my services who wanted creative thinking oftener THAN ONCE A DAY, SHOULD RELIEVE ME OF MY ADMINISTRATIVE WORK, otherwise I might describe myself as a hard worker during the day on the mechanics of the job and a creative thinker at night on my own time."

• **Conditioned thinking** can prevent a person from adopting the new perspective that may be needed to solve a problem. A common response in the business world is to ask employees to "think outside the box". In contrast, the zoo mammal, when moved to a larger cage, continues to pace an area similar to that of the old cage [Biondi, 1980].

Beveridge [1955] suggests several ways to break free from conditioned thinking. Set the problem aside for a while then resume; as discussed in the previous section, temporary abandonment helps by allowing the established thought pattern to fade, perhaps permitting a new one to start. More drastically, one may need to start over from the beginning with a very different approach. Talking over the puzzle with others or writing up the project can provide the new perspective. Reading related papers, or even potentially relevant papers on different subjects, at least will drive a wedge between conditioned thinking and the problem. They also may evoke a useful analogy. The value of abandoning conditioned thinking is the lesson of the Sufi story, The River, which began this chapter.

* * *

More dangerous than the factors preventing insight is excessive confidence in one's insight. An almost universal characteristic of insight is the *conviction of truth*. Unlike the scientist's normal attitude that hypotheses can be disproved but not proved, the flash of insight often is accompanied by a certainty that the discovered pattern is so elegant that it must be true. This certainty is a scientific pitfall that can undermine the objective undertaking of the fourth stage of research: verification. When Platt and Baker [1931] polled scientists and asked whether they had ever had a revelation that turned out to be wrong, only 7% said that their insights were always correct.

Part of this conviction of truth may be attributable to the sudden breakthrough of pattern recognition. The greater the breadth of the pattern and its apparent ability to account for disparate observations, the greater the conviction of truth. Yet cold analysis may reveal fatal flaws in the insight.

> Last winter my scientist wife and I talked often about her unexplained research results. She tested and rejected many hypotheses. Then I had an exhilarating insight into what the 'true' explanation was. I explained my complex model to her, as well as the surprising and therefore diagnostic results that my model predicted for two experiments that she had not done yet. She pointed out that the model was contrary to conventional theory; I agreed with a smile and with unshaken conviction of my

model's accuracy. Although she was dubious of the model, she undertook the two experiments. One result fit my predictions and one contradicted them, and today only a dim echo of my model is accepted by either of us. Yet my exhilaration at discovering the model was not balanced by a corresponding disappointment at seeing the model proved wrong, perhaps because my emotional involvement with the problem was an intrigued outsider's interest rather than an anxiety of frustrated scientific progress.

About a month after my model failed, my wife was continuing the experiments at another lab and called me to say: "We've always been thinking of the energy barriers as peaks. What if they are troughs instead?" Immediately I felt that she was right, that she had solved the problem. Her answer was so much simpler than mine had been. With this new perspective, we were amazed that we all had been obtuse for so long. Of course, not everyone is as certain as we are that she is right.

"When you have at last arrived at certainty, your joy is one of the greatest that can be felt by a human soul." [Pasteur, 1822-1895,b].

* * *

The Royal Way

Success or failure in reaching insight often depends on the path followed. Once the goal is achieved, however, the path becomes irrelevant to the evaluation of that insight.

"But any pride I might have felt in my conclusions was perceptibly lessened by the fact that I knew that the solution of these problems had almost always come to me as the gradual generalization of favourable examples, by a series of fortunate conjectures, after many errors. I am fain to compare myself with a wanderer on the mountains, who, not knowing the path, climbs slowly and painfully upwards, and often has to retrace his steps because he can go no farther -- then, whether by taking thought or from luck, discovers a new track that leads him on a little, till at length when he reaches the summit he finds to his shame that there is a royal way, by which he might have ascended, had he only had the wits to find the right approach to it. In my works, I naturally said nothing about my mistakes to the reader, but only described the made track by which he may now reach the same heights without difficulty." [Helmholtz, 1891]

Between 1899 and 1904 French mathematician Henri Poincaré considered many of the same factors, including in 1904 the same term 'principle of relativity', that Albert Einstein brought together in his 1905 paper on special relativity. Yet Poincaré was unable to reach the same insight first. Poincaré says in his 1911 letter of reference for Albert Einstein:

"I do not mean to say that all these predictions [by Einstein] will meet the test of experiment when such tests become possible. Since he seeks in all directions, one must, on the contrary, expect the majority of the paths on which he embarks to be blind alleys. But one must hope at the same time that one of these directions he has indicated may be the right one, and that is enough. This is exactly how one should proceed. The role of mathematical physics is to ask questions and only experiment can answer them." [cited by Hoffmann, 1972]

Concerning his 1915 discovery of general relativity, Albert Einstein [1879-1955] said:

> "In the light of knowledge attained, the happy achievement seems almost a matter of course, and any intelligent student can grasp it without too much trouble. But the years of anxious searching in the dark, with their intense longing, their alternations of confidence and exhaustion, and the final emergence into the light -- only those who have themselves experienced it can understand that."

Helmholtz would have understood it.

* * *

How Does Insight Work?

Insight is the least controllable aspect of scientific research. It can be encouraged, however, by immersion in an examination of all relevant evidence, followed by relaxation and temporary abandonment of the problem. Furthermore, we know that conditions such as interruptions can prevent insight. But what is the mechanism of insight? What marriage between data and pattern recognition is performed in the brain, resulting in the birth of insight? I don't know, but I think we have seen some clues.

J.E. Teeple, a respondent to Platt and Baker's [1931] questionnaire, may have hit upon the most decisive element, concentration:

> "It is this deep concentration that is the most valuable asset in the solution of any problem. We speak of thinking and try to divide it into conscious, subconscious, and completely unconscious, which I think is an error. In deep concentration on any subject you are not only unconscious that you are thinking but you are unconscious of everything else around you."

Imagine substituting the word 'concentration' for 'insight' in the previous sections on conditions favoring insight and obstacles to insight; usually the discussions still would be valid. It appears that concentration is a necessary but not sufficient condition for insight.

Another clue to the mechanism of insight may come from the relationships of data and hypothesis generation to insight. We usually feel that there are too few data to force a conclusion, or more likely not enough data of the needed type. In contrast, the geometric expansion of science in this century often creates the converse problem: there are too many data of too many relevant but somewhat different types to grasp and consider simultaneously. One is left with the vague hunch that the answer is hidden somewhere in the masses of data; perhaps, some filter or new perspective is needed to extract the key observations and their relationships.

Do we achieve insight through subconscious processing of all possible permutations of the evidence, or of only a subset? Consider the following two contrasting viewpoints on hypothesis generation:

> "Mathematical creation does not consist in making new combinations with mathematical entities already known. Anyone could do that, but the combinations so made would be infinite in number and most of them absolutely without interest. . . The true work of the inventor consists in choosing among these combinations so as to eliminate the useless ones, or rather to avoid the trouble of making them." [Poincaré, 1905]

> "The effort to solve a problem mentally is a constant series of trials and errors. The mind in searching for a solution considers in rapid succession a long series of conceivable answers, each of which is almost instantly rejected on account of some

> obvious objection. Finally in this process of trial and rejection we more or less accidentally stumble upon an answer to which the objection is not so obvious. The smooth course of trial and rejection is brought to a halt. Our attention is arrested." [Platt and Baker, 1931]

> "Discovery is something a computer (if constructed and programmed well enough) could do, and do as well (even better) than any human who ever lived." [Jason, 1989]

Insight spans, I suspect, a continuum from conscious to unconscious. Platt and Baker [1931] may be partly right, *if* much of the filtering of ideas occurs either subconsciously, on the fringe of consciousness, or in such a brief conscious flash that we are barely aware of it. The less absurd ideas require a little more conscious focus before they can be discarded. Yet Poincaré grasps a central point, missed by Platt and Baker, that the successful scientist owes as much to excluding broad regions from trial-and-error evaluation as to the evaluation itself. A better-trained computer is *not* the solution.

The approach of considering all possible data permutations is hopeless. What is needed is a leap of insight to the crux. As in chess, the best player does not simply examine all permutations methodically; instead the master visualizes patterns and focuses in on a small subset of the possibilities. I think that the pitfall of conditioned thinking offers a useful perspective: rather than systematically scanning a huge number of possible explanations, scientific thoughts get trapped among a few patterns, like a song that you cannot get out of your head. Relaxation and temporary abandonment may work because the problem continues to pop unbidden into the fringe of consciousness, interspersed with seemingly unrelated thoughts, until suddenly the mind sees the problem in the context of a potentially explanatory pattern.

From a neurobiologist's perspective, the brain has a vast number of schemata-templates [Calvin, 1986]. Each schema is a neural pattern or pathway, formed and reinforced by electrical current flow. Each schemata-template is triggered whenever we see or experience something that seems to fit the pattern. Boyd [1985] describes hypothesis creation as "finding new combinations of previously understood ideas and concepts." If schemata are reinforced via electrical current flow in the brain, could insight be a sudden flow in parallel of schemata that had never previously flowed simultaneously?

> "It is a wondrous thing to have the random facts in one's head suddenly fall into the slots of an orderly framework. It is like an explosion inside. . . I think that I spend half my time just talking and listening to people from many fields, searching together for how [plate tectonics] might all fit together. And when something does fall into place, there is that mental explosion and the wondrous excitement. I think the human brain must love order." [marine geologist Tanya Atwater in 1981, during the period in which the new paradigm of plate tectonics was revolutionizing geology; cited by Calvin, 1986]

Even if the idea of insight as schema generation is correct, its practical usefulness may be small. It does, however, reveal a potential problem: those insights are favored that are similar to ideas and concepts that are already established. Breakthroughs to a radically new perspective are not fostered. Contrast these incremental advances with the following:

> "In each of the 1905 papers, Einstein has totally transcended the Machian view that scientific theory is simply the 'economical description of observed facts.' None of these theories, strictly speaking, begins with 'observed facts'. Rather, the theory tells us what we should expect to observe." [Bernstein, 1982]

* * *

Alternative Paths to Insight

The preceding sections give the misleading impression that insight only follows a prolonged search for the explanation to one's observations. Other sources of insight, however, can be just as fruitful. Chance can play a prominent role in discovery. Many breakthroughs are by amateurs, or at least by those with scanty experience of the relevant evidence. And one of the most powerful ways of achieving an insight is to borrow from another field.

Unexpected Results

Chance makes an influential, yet often overlooked, contribution to discovery. For example, Roentgen's discovery of X-rays began with an accident: photographic plates left near a discharge tube were inexplicably blackened. When Alexander Fleming noticed a strange mold growing on his culture dish, he isolated it, purified it, and discovered penicillin, the first antibiotic.

Rather than providing a variation on existing themes, chance discovery can lead to a totally new perspective. One can seek insight but cannot seek chance discovery. One can, however, open oneself to this type of discovery [Beveridge, 1955]:

> Be alert for any unexpected result. Resist the temptation to rationalize away or discard them. Remember that observations that do not fit predictions, though often ignored, sometimes are responsible for the new paradigm. "Remain alert and sensitive for the unexpected while watching for the expected" [Beveridge, 1955]. Try novel procedures, to increase the likelihood of encountering surprises.

In seeking insight from unexpected results, we may encounter pitfalls instead. It is easy to become distracted and pulled in a new direction by every unexpected result, so that one seldom completes a suite of experiments. This pitfall is often avoidable: simply flag the unexpected data and come back to them later. One can easily confuse the 'chance-in-a-lifetime' result with trivial results, failing to follow up on the former or wasting considerable time on the latter. Louis Pasteur [1822-1895, a] repeatedly said, "In the field of experimentation, chance favors only the prepared mind." One needs considerable background, to recognize the unexpected result and to evaluate correctly its importance and significance.

Transfer From Other Disciplines

One path to insight that is frequently successful, yet underutilized, is the extension of a technique, algorithm, relationship, or equipment from one field to another:

> "For every original discovery there are dozens of important advances which are made simply by recognizing that a scheme developed for one field or application can be applied to another." [Wilson, 1952]
>
> "Making variations on a theme is really the crux of creativity." [Hofstadter, 1985]

For example, Einstein's emphasis on reference frames was a key not only to relativity but also, much later, to the paradigm of plate tectonics in geology. The recent physics paradigm of chaos is creating breakthroughs in oceanography, meteorology, biology, and earthquake mechanics. These

examples are of theoretical concepts, but even more consistently productive is the application of new techniques and instruments to empirical research.

Many discoveries are driven by technology, by the sudden ability to make a new type of measurement or to make a much more accurate measurement than before. The inventors of the electron microscope, laser, and CT scan could never have imagined the realms that these devices would explore. Recognition of the applicability of a new instrument or technique is easiest for the person who knows the problem, not for the person who develops the technique. One does not necessarily even have to match the new technique with a specific problem. Discovery can result just from a 'fishing expedition' with a new kind of observation of an old phenomenon.

Because science is increasingly specialized, researchers seldom are aware of technical or conceptual developments in 'unrelated' fields. Thus a potential application may go unrecognized for many years. This approach also is underutilized because it is haphazard, almost always stumbled upon rather than deliberately sought out. Yet it can be fostered. Seeking new applications to one's field is a strong incentive for reading outside of one's field. Such goal-oriented reading can be extremely productive.

Breakthroughs by Amateurs: the Outsider Perspective

Breakthroughs by 'amateurs' are a phenomenon that seems to run counter to the philosophy of acquiring all relevant data to assist in reaching an insight. Actually, the 'amateur' usually is not a scientific novice, but an experienced scientist who has just changed fields. The neophyte brings to a new field the established disciplines of scientific method but not the prevailing assumptions or prejudices of the entrenched leadership of that field. The newcomer may also bring a technique or concept to a different field, as mentioned above.

A related phenomenon is breakthroughs by young scientists: most revolutions within each field of science are led by the younger generation (in physics, for example, by those under 30 years old). This generation may have more energy than the older generation, but it also has less efficiency and much less knowledge. The higher discovery rate among relative newcomers to a field stems from their greater flexibility of thought, due to less ingrained assumptions and conclusions.

Published errors -- whether in assumptions, data, or interpretations -- are stumbling blocks to further insights, particularly for researchers who have long accepted them. It does not follow, though I have heard the argument made, that it is better for one to avoid reading intensively in one's specialty. Oliver [1991] points out that seeking breakthroughs outside one's specialty brings hazards as well as opportunities. For example, physicist Lord Kelvin calculated the age of the earth and was dogmatic that his result was correct. He ignored geologists' evidence for a much older age, partly because the evidence was not from his field.

Knowing all the relevant evidence is essential, but equally essential is alertness to the basis for one's assumptions and interpretations, and critical reevaluation of that basis.

Changing fields is a drastic means of chasing insights. Changing subjects within the same field can unleash creativity [Loehle, 1990]. Another pragmatic alternative is simply to try out different perspectives. The following problem illustrates this approach.

> A gnat flies deep into your ear and repeatedly collides with your eardrum. How can you solve the problem?
>
> First, consider the consequences of initial failure on future trials (Chapter 5): squirting water into your ear might wash the gnat out, but a dead gnat may be even harder to extract than a live gnat. Instead, try the problem-solving technique of

> choosing the perspective of others -- in this case, the gnat's perspective. Does the gnat want to be in or out? Is there anything that you can do to influence the gnat's behavior? Gnats, like most flying insects, are attracted to light. So shine a light in your ear, and help the gnat escape.

For those who are more committed to attaining insights than to persisting with their current research projects, consideration of the following questions may suggest more fruitful research directions [Oliver, 1991]. Is the present discipline becoming more isolated and losing touch with the rest of the field? What is its past and present rate of progress and discoveries, and are they accelerating or decelerating? Are many datasets still not understood, thereby indicating the potential for new insights? Is the field undergoing long-term changes, such as from observational to theoretical or perhaps toward increasing interaction with some other discipline? Do gaps between specialties create opportunities for dual-specialty science, and will their exploitation require intensive background study or selection of co-workers in that specialty? Is there an alternative research thrust to that followed by the majority of researchers in the field?

* * *

From Puzzle Solving . . .

The magnitude of the creative leap forms a continuum, from minor problem solving to major creative insight to mystical experience. The thrill of major creative insight or mystical experience is quite rare, yet most scientists capture a taste of that thrill every day in the small-scale problem solving that is a characteristic part of science. Indeed, many scientists have puzzle-solving hobbies such as chess, bridge, and reading mysteries -- hobbies that further gratify the craving for insights of any size.

Some classic puzzle-solving techniques also foster both insight and scientific problem-solving:

- redefine the problem by breaking it down into several components, then attack one or more of these pieces individually;
- decide which thread to grasp, to start unraveling the puzzle;
- analyze all assumptions and detect inappropriate, overlooked, or invalid assumptions;
- provisionally assume an answer, then look at its implications for the problem.

Perhaps a hobby of puzzle solving can improve our ability to recognize hidden scientific assumptions. Killeffer [1969], among others, suggests that the practice of puzzle solving improves the ability of the mind to see patterns and associations. This ability, like other acquired skills, can be enhanced by practice.

> "It began with little things, certain small clinical changes which I observed. Little things can be important. Even more important is the ability -- call it knack, hunch, providence, good luck, whatever -- to know what you are looking for and to put two and two together. A great scientist once said that genius consists not in making great discoveries but in seeing the connection between small discoveries. . . Small disconnected facts, if you take note of them, have a way of becoming connected." [Percy, 1987]

> "It's just like doing a jigsaw puzzle: whenever you think that there is no piece that can possibly fill a blank space, you don't just throw up your hands and insist that

only a miracle will solve the problem. You keep looking, and eventually you find something that links together the parts of the puzzle." [Calvin, 1986]

But in science, unlike in puzzle solving, the problem may be impossible to solve. That uncertainty is part of the challenge.

* * *

. . . to Mystical Experience

Many scientists will be bothered or even offended by my inclusion of mystical experiences on the continuum of insight intensity. Mystical experiences are automatically lumped with emotions and other non-rational and therefore non-scientific subjects. Perhaps psychologist Abraham Maslow's term 'peak experience' or James Joyce's term 'epiphany' is more palatable. I argue that major creative insight is a goal-oriented subset of mystical experience rather than a fundamentally different phenomenon. But I am content if the reader agrees that the two phenomena exhibit some surprisingly strong parallels, as illustrated by the following examples.

St. Augustine [354-430 A.D., b] described a mystical experience as "a moment of supreme exaltation, followed by gradual absorption back into the normal state, but with resulting invigoration and clearer perception."

"Then followed months of intense thought, in order to find out what all the bewildering chaos of scattered observations meant, until one day, all of a sudden, the whole became clear and comprehensible, as if it were illuminated with a flash of light. . . There are not many joys in human life equal to the joy of the sudden birth of a generalization illuminating the mind after a long period of patient research." [Kropotkin, 1899]

Naturalist Annie Dillard read accounts of people blind since birth who were suddenly given sight by cataract surgery, and of their remarkably variable and sometimes frightened reactions to this new vision. She was particularly captivated by the experience of one young girl who stared, astonished, at what she finally recognized as a tree. This was not a tree such as you or I have ever seen, but "the tree with the lights in it" [Dillard, 1974]:

"When her doctor took her bandages off and led her into the garden, the girl who was no longer blind saw 'the tree with the lights in it.' It was for this tree I searched through the peach orchards of summer, in the forests of fall and down winter and spring for years. Then one day I was walking along Tinker Creek thinking of nothing at all and I saw the tree with the lights in it. I saw the backyard cedar where the mourning doves roost charged and transfigured, each cell buzzing with flame. I stood on the grass with the lights in it, grass that was wholly fire, utterly focused and utterly dreamed. It was less like seeing than like being for the first time seen, knocked breathless by a powerful glance. The flood of fire abated, but I'm still spending the power. Gradually the lights went out in the cedar, the colors died, the cells unflamed and disappeared. I was still ringing. I had been my whole life a bell, and never knew it until at that moment I was lifted and struck. I have since only very rarely seen the tree with the lights in it. The vision comes and goes, mostly goes, but I live for it, for the moment when the mountains open and a new light roars in spate through the crack, and the mountains slam."

"Thus my mind, wholly rapt, was gazing fixed, motionless, and intent, and ever with gazing grew enkindled, in that Light, . . . for my vision almost wholly departs,

> while the sweetness that was born of it yet distills within my heart." [Dante Alighieri, 1313-1321]
>
> "The flame of conception seems to flare and go out, leaving a man shaken, and at once happy and afraid. . ." [John Steinbeck, 1954; cited by Calvin, 1986]

* * *

Living science, for me, is punctuated illumination, less blinding but more frequent than the experiences of Dillard and Dante. I too am struck and go on ringing, and I am so addicted that neither the countless minor frustrations nor occasional stagnations can fully damp the memory and obsession with this ringing. Perhaps someday I will pick a problem that defies solution, search for years without finding, and finally claim that I really became a scientist because of motivations other than the thrill of insight. Or perhaps I will see "the tree with the lights in it." For now, I go on ringing.

> "I feel like shouting 'Eureka!', awakening the camp. But caution reasserting itself, I satisfy myself with a broad smile instead, and look overhead at the drifting clouds. I must try this out, see just how much of the universe's known mechanism can be appreciated from this new viewpoint." [Calvin, 1986]

Chapter 9: The Scientist's World

Them/us – one of the simplest and potentially most devastating human classifications – is the topic of this chapter. Here we examine our relationships as scientists to various other groups, within and outside science.

Scientist and Lay Person

In what way -- if any -- are scientists unique? In the following passages, the difference between scientists and other people is described by Herbert Spencer, Annie Dillard, and William Shakespeare (though scientists were not the subject of Shakespeare's thoughts). Yet Spencer's perspective is laced with arrogance, Dillard's with apparent envy, and Shakespeare's with joy.

> "Is it not, indeed, an absurd and almost a sacrilegious belief that the more a man studies Nature the less he reveres it? Think you that a drop of water, which to the vulgar eye is but a drop of water, loses anything in the eye of the physicist who knows that its elements are held together by a force which, if suddenly liberated, would produce a flash of lightning? . . . Think you that the rounded rock marked with parallel scratches calls up as much poetry in an ignorant mind as in the mind of a geologist, who knows that over this rock a glacier slid a million years ago? The truth is, that those who have never entered upon scientific pursuits know not a tithe of the poetry by which they are surrounded. Whoever has not in youth collected plants and insects, knows not half the halo of interest which lanes and hedgerows can assume. Whoever has not sought for fossils, has little idea of the poetical associations that surround the places where imbedded treasures were found. Whoever at the seaside has not had a microscope and aquarium, has yet to learn what the highest pleasures of the seaside are. Sad, indeed, is it to see how men occupy themselves with trivialities, and are indifferent to the grandest phenomena -- care not to understand the architecture of the Heavens, but are deeply interested in some contemptible controversy about the intrigues of Mary Queen of Scots!" [Spencer, 1883]

> "I cherish mental images I have of three perfectly happy people. One collects stones. Another -- an Englishman, say -- watches clouds. The third lives on a coast and collects drops of seawater which he examines microscopically and mounts. But I don't see what the specialist sees, and so I cut myself off, not only from the total picture, but from the various forms of happiness." [Dillard, 1974]

> "And this our life exempt from public haunt
> Finds tongues in trees, books in the running brooks,
> Sermons in stones and good in every thing.
> I would not change it."
> [Shakespeare, 1600]

Many lay people hold a stereotypical view of scientists. We are perceived to be:

- very intelligent;
- myopic in interest, focusing on precise measurements of a tiny subject;
- objective in both measurements and interpretations;
- conservative, accepting no interpretation or conclusion unless it has been proved beyond doubt;

• oblivious of the possibly harmful applications of our research results; and

• above all, completely rational and unemotional.

These perceptions are, in part, responsible for the authority of science. Like all stereotypes, however, they depersonalize. Scientists are above average in intelligence, and I have known individual scientists who were myopic, precise, conservative, or oblivious. I have seen scanty evidence, however, that scientists in general fulfill the stereotypes above. Only our publications are completely rational and unemotional; their authors, in contrast, are passionate.

Scientists do tend to differ from most lay people in their techniques, particularly in their embracing of the scientific methods. But of course every kind of specialist differs from lay people in embracing certain techniques and achieving professionalism in exercising those techniques. Like many other specialists, scientists inadvertently build a barrier of jargon. The jargon permits efficient, exact communication among specialists but seems to the outsider to be deliberately exclusive and abstruse. The motivations of scientists -- to the extent that one can generalize -- resemble those of artists; they differ only in degree from most other people.

We are craftsmen, not geniuses.

* * *

Science and Society

> On seeing the culmination of the Manhattan Project (the first detonation of a nuclear bomb) J. Robert Oppenheimer [1945] quoted from the Bhagavad Gita: "I am become Death, the shatterer of worlds."

Some species are solitary and some are social. People try to gain the advantages of both strategies, living together in an interdependent society but encouraging individuality. Inevitably conflict erupts between individual and societal needs. This balancing act is acutely felt by scientists, who accept support but not control from society. The scientist listens to cultural guidelines but personally selects values and priorities [Campbell, 1988b]. The "age-old conflict between intellectual [or moral] leadership and civil authority" [Bronowski, 1973] was fought by Socrates, Jesus, Galileo, Darwin, and Gandhi, as well as by scientists whose names are forgotten. Einstein [1879-1955] may have underestimated the strength of the opposition in his 1953 comment:

> "In the realm of the seekers after truth there is no human authority. Whoever attempts to play the magistrate there founders on the laughter of the Gods."

Scientific responsibility is personal:

> In 1933 Leo Szilard was stopped at a red light while walking to work, when suddenly he realized that neutron bombardment could potentially initiate an explosive chain reaction. He faced the choice of keeping his discovery secret or publishing it, of delaying its use or allowing its abuse. Seeking secrecy, he took out a patent and assigned it to the British admiralty [Bronowski, 1973], but of course development of the atomic bomb would not be slowed by a patent. In 1939 he ghost-wrote a letter, signed by Einstein, which warned President Roosevelt of the danger of nuclear weapons.

Szilard would have empathized with the anonymous statement [cited by Matthiessen, 1978]: "God offers man the choice between repose and truth: he cannot have both." Then, as now, applied

science was not confined to discovering what technologies are possible; it also predicted consequences and side effects of those technologies.

About 4% of the U.S. population has a degree in science or engineering. For most of the others, exposure to science is generally indirect: basic science ⇒ applied science ⇒ engineering ⇒ technology [Derry, 1999]. Technology is the tangible result of combining applied science with engineering and business skills.

Popular opinion of science and scientists waxes and wanes with attitudes toward technology. After the technological enthusiasm and optimism of the sixties, the rock group Jefferson Starship [1970] sang: "Do you know we could go, we are free. Anyplace you can think of, we could be." A decade later, however, a society that seldom can think more than four years ahead encountered the consequences of past technological decisions and found that the technological 'gift' of comfort actually has a price. "Comfort, that invader that enters as a visitor, stays as a guest, and becomes master" (Sufi saying). Someone must be blamed, and a musician said to my wife: "Oh, you're a physicist. I suppose you build bombs." Mea culpa, mea maxima culpa. In the nineties, technological development led to improved standards of living and an exuberant tech bubble. Ethical concerns and fears about technological developments have shifted from atomic weapons to genetic engineering.

> "To every man is given the key to the gates of heaven; the same key opens the gates of hell." [Buddhist proverb, cited by Feynman, 1988]

> "We fear the cold and the things we do not understand. But most of all we fear the doings of the heedless ones among ourselves." [a shaman of the Arctic Inuit, cited by Calvin, 1986]

The beneficiaries of technology have the opportunity to see its shortcomings. In contrast, people whom I have met in underdeveloped countries simply hunger for its rewards and for its escape from boring drudgery. Few of the critics of science accuse it of being evil, but many accuse it of being amoral. One can counter such arguments by asking whether the professions of farming and carpentry are also guilty of amorality. Or one can recall that science's highest value is truth (Bronowski, 1978), and that we judge truth from criteria of beauty, simplicity and elegance; is this amorality? But such arguments miss the point. Some people simply are becoming disillusioned with technology, and they are replacing the illusion of technology as magic bullet with one of technology as evil destroyer.

> "Daedalus, who can be thought of as the master technician of most ancient Greece, put the wings he had made on his son Icarus, so that he might fly out of and escape from the Cretan labyrinth which he himself had invented. . . He watched his son become ecstatic and fly too high. The wax melted, and the boy fell into the sea. For some reason, people talk more about Icarus than about Daedalus, as though the wings themselves had been responsible for the young astronaut's fall. But that is no case against industry and science. Poor Icarus fell into the water -- but Daedalus, who flew the middle way, succeeded in getting to the other shore." [Campbell, 1988b]

* * *

The relationship between science and society is changing, in response not only to evolving perceptions by society, but also to other evolutionary pressures. Both the tasks and needs for science are adapting accordingly.

Science has transformed the highly generalized and adaptable human species into the most adaptable species that the earth has ever seen (Bronowski, 1978). Yet arguably we have increased

our need for adaptability at an even faster pace, because each technological change can have unforeseen interactions, either with the environment or with other technological changes. In response, many scientists are becoming environmental and technological troubleshooters.

Biological evolution demonstrates that specialization only survives in a static environment. Society's needs concerning specialization versus adaptability are changing: the pace of technological change is increasing, professions are waxing and waning, and therefore our society needs individuals with the ability to move into newly emerging careers. We also need individuals comfortable in interdisciplinary teams.

Scientific education is evolving in response to these changes. For graduate study, the change is less than one might expect: graduate programs entail specialized research, but the competencies learned actually increase the student's adaptability. The old notion of an early academic education followed by a lifetime profession may be obsolete; it is certainly incomplete. The rapid pace of scientific and technological change means that knowledge is not static and education is never really finished. Increasingly, the educational system is being used for retooling and redirection. Students are teaching the professors by communicating the perspectives and needs of industry. Conversely, the students are taking practical applications of their course work to the work-place *immediately*, not years later.

* * *

Major changes of any kind are stressful -- to individuals, groups, and society. The redirection of scientific efforts and education, in response to societal needs, is non-trivial, emotionally taxing, but essential.

The public and politician, having grown up with textbook-science facts, expect certainty from scientists. We, in contrast, savor the uncertainty implicit in forefront science, where ideas are explored, modified, and usually discarded. We offer the authority of science with humility. More than once in the history of science, scientists have had to fight for the privilege of questioning authority. This popular expectation of scientific certainty creates roadblocks, when the implications of scientific research are that society needs to take expensive action. Scientific debate provides a political excuse for societal inaction, even if the key issues are agreed upon among scientists.

> An example is the greenhouse effect, concisely summarized by Stevens [1992a]. Researchers agree that: (1) atmospheric carbon dioxide is rising due to burning fossil fuels and clearing rainforests, (2) atmospheric carbon dioxide will have doubled within the next 60 years, (3) increased carbon dioxide warms the earth through the greenhouse effect, and (4) as a consequence, the earth will warm up during the coming decades.
>
> Some issues are still being debated: How much greenhouse warming has already occurred? How fast and how much warming will the doubling of carbon dioxide induce? What will the local climate effects be? Uncertainty over these questions obscures consensus on the former concerns. We postpone remediation; 'wait-and-see' is cheaper.

Technological innovations are the most frequent and obvious contributions of science to society, but occasionally science has a more fundamental impact: it can change humanity's self-image [Derry, 1999], by generating "the light which has served to illuminate man's place in the universe" [J.F. Kennedy, 1963]. The determinism of Newton's mechanics and the indeterminacy of quantum mechanics challenge our assumption of free will, but this assumption is rooted too firmly to be damaged. The Copernican revolution did not merely overthrow the concept of Earth as center of the rotating universe; it dislodged humanity also from that position. Darwin's theory of biological

evolution by natural selection forced another radical revision of self-image: not people as the designated masters of animals, but people as distant relatives of all other animals. The Copernican revolution was resisted and the Darwinian revolution is still resisted because of unwillingness to relinquish self-importance.

> "Most laymen, when they contemplate the effect physics may have had upon their lives, think of technology, war, automation. What they usually do not consider is the effect of science upon their way of reasoning." [Baker, 1970]

* * *

Science and the Arts

As scientists reach out to society, attempting to dispel misconceptions of science, shall we consider the arts as allies or opponents? Are there two cultures, scientific and literary, separated by a gulf of misunderstanding and conflicting values? C. P. Snow [1964] argued persuasively that there are. Most of us have met both scientists and artists whose scorn for the other culture is vast:

> "It may be important to great thinkers to examine the world, to explain and despise it. But I think it is only important to love the world, not to despise it, . . . to regard the world and ourselves and all beings with love, admiration and respect." [Hesse, 1923]

> "In fact, pure science . . . is at once a substitute for logic as a discipline for the mind and an expression of an insatiable desire for the conquest of all knowledge, for an intellectual mastery of the universe." [Burns, 1963]

> "The highest Art of every kind is based upon Science – that without Science there can be neither perfect production nor full appreciation." [Spencer, 1883]

Such individuals separate themselves from a potentially enriching aspect of life by a barrier built at least partially upon misconceptions. The barrier is permeable: many scientists, particularly physicists, are also amateur musicians. A few remarkable individuals, such as Leonardo da Vinci and Benjamin Franklin, excelled in both cultures. I suspect that today's cultural separation is largely a failure to communicate.

Science and art share some key features. Creativity is the source of vitality in both. Science has no monopoly on subjecting that creativity to a rigorous, critical attitude, as any art critic would point out. Virtuosity of both design and technical performance is a hallmark of the best in science and the arts. Both science and poetry are "acts of imagination grounded in reality. . . These two great ways of seeing lie on the same imaginative continuum." [Timpane, 1991].

The craftsmen differ more in their tools than in their skills.

* * *

Science and Pseudoscience

In examining links between science and society, or between science and art, we assume agreement at least on what science is and is not. But how does one distinguish science from pseudoscience? Most scientists do so on a case-by-case basis, with a demarcation that is subjective and value-dependent.

The prevailing discriminator is use of the 'scientific method': sciences all use the scientific method, and pseudosciences either do not use it, misapply it, or minimize a crucial portion of it. The problem with this criterion, however, is its invalid premise -- that a single scientific method is used by all sciences. This book is based on a different premise: the sciences share a suite of scientific methods, but the emphasis on individual techniques varies among and within sciences.

This revised discriminator -- use of the suite of scientific methods -- is employed by scientists with reasonable success. Astrology, UFO's, and psychic healing are considered by many to be pseudosciences, because they lack a well-controlled observation base. Parapsychology, in contrast, is very rigorous experimentally, yet most scientists reject it because of inadequate replicatability and because its results challenge their key assumptions (e.g., can the outcome of an experiment be affected by the experimenter's wishes?). Immanuel Velikovsky's [1967, 1977] ideas about colliding planets are rejected in spite of his volumes of supporting evidence, because of his complete absence of objectivity in evidence evaluation.

Are political science and sociology really sciences? For many scientists, the answer to that question depends less on each field's methods than on respect for their results. That decision should be based on reading the original literature or at least textbooks, rather than on such 'data' as newspaper editorials.

* * *

The challenge of separating science from pseudoscience has intrigued many philosophers of science. This goal inspired the birth of falsificationism, Karl Popper's philosophy that science should concentrate on trying to falsify hypotheses (Chapter 7). Popper was uncomfortable with the 'scientific' theories of Marx, Freud, and Adler, particularly in the way these theories seemed to account for *any* observation:

> "What, I asked myself, did it confirm? No more than that a case could be interpreted in the light of the theory. But this meant very little, I reflected, since every conceivable case could be interpreted in the light of Adler's theory, or equally of Freud's. . . It was precisely this fact -- that they always fitted, that they were always confirmed -- which in the eyes of their admirers constituted the strongest argument in favor of these theories. It began to dawn on me that this apparent strength was in fact their weakness." [Popper, 1963]

The line that Popper found to separate science from pseudoscience was the criterion of falsifiability: "statements or systems of statements, in order to be ranked as scientific, must be capable of conflicting with possible, or conceivable, observations." Hypothesis testing certainly is an integral component of all sciences.

Thomas Kuhn used a quite different discriminator: every science has a ruling paradigm that explains a wide variety of observations and guides research, whereas fields that lack such a paradigm are in the 'pre-science' stage. Kuhn's discrimination of pre-paradigm and paradigm-guided research is useful (Chapter 7), but it does not follow that pre-paradigm fields are pre-science or pseudoscience. For example, sociology and parts of psychology lack consensus on a unifying paradigm, but that lack does not constitute grounds for rejecting their findings.

* * *

Why have scientists largely ignored the efforts to identify a science/pseudoscience demarcation? They reject the premise of this quest: "I define the criterion that determines whether you are a scientist or pseudoscientist." The label of 'pseudoscience' accomplishes more harm than good.

This pejorative term substitutes for a rational discussion of the scientific strengths and weaknesses of fields. The result is ostracism rather than inducement for a field to respond constructively to outside criticism.

* * *

Applied and Basic Research

The bridge between science and society is the teamwork of basic and applied research. Yet sometimes it seems that the distance across that bridge is too great for clear perception.

Conflict: Applied vs. Basic Research

That some non-scientists hold distorted views of science and technology is not surprising. Even scientists sometimes succumb to stereotypes concerning science, particularly regarding applied vs. basic research.

The choice between applied and basic research is a watershed career decision. Perhaps it is to be expected that the individual will reinforce that choice, by emphasizing the perceived disadvantages of the rejected option.

My own perspective of the dichotomy between basic and applied research is from the physical sciences. The boundary is fuzzier and perhaps the prejudices are fewer in the social sciences, because study of behavior is implicitly alert to human applications. I have worked primarily in basic research, and I have often heard the academics' stereotypes about industry scientists ('materialistic', 'less intelligent', 'less creative'). This prejudice is particularly obvious in the academic's use of the term 'pure research' to describe basic research, as if applied research is somehow impure. Yet I also worked for several years in industry, where I saw corresponding stereotypes by industry researchers toward academics ('ivory tower', 'dilettantes and dabblers', 'groundless pomposity'). Both sets of stereotypes had more to do with personal ego massage than with real differences. Some generalizations are possible, if we are mindful of frequent exceptions.

[Harris, 1970]

The methods of basic research and applied research are the same.

Basic research seeks knowledge of any kind. Applied research is alert to and partially directed by potential practical applications. This distinction is not absolute, however. Branches of basic research with obvious implications for society are more fundable than other basic research. An applied researcher may devote prolonged effort to basic issues if they have been inadequately

developed by academics. Indeed, some industrial analysts attribute Japanese technological success partly to the willingness of Japanese industry to establish a firm theoretical foundation. Thus applied research does not merely follow up on basic research; the converse can be true.

Some basic researchers claim that they are free to explore the implications of unexpected results, whereas applied researchers are compelled to focus on a known objective. Yet both pursue applications of their discoveries, whether industrial or scientific, and both allow potentially fruitful surprises to refocus their research direction.

Successful industrial competition means not only getting ahead in some areas, but also keeping up in others. Often, it is more efficient for a company to introduce and apply published work by others than to initiate experiments. Applied researchers may experience conflict between their scientific value of open communication and the business need for confidentiality. Applied researchers tend to be more alert than basic researchers to potential applications for their research of discoveries in a different field. Applied research is generally more mindful of economic factors, more cognizant that an approach may be theoretically viable yet financially or otherwise impractical.

Usually the academic researcher can maintain the illusion of having no boss, whereas the chain of command in industry is obvious. It may be easier to start a pilot project in industry. Go/no-go decisions are more frequent too; continuation of the project must be defended at every step.

Some applied researchers [e.g., Killeffer, 1969] see academic research as a 'stroll through the park,' with no pressure to produce or to work efficiently. Job security in either type of research affects productivity pressure; probably the most pressured are researchers on 'soft money' -- dependent on funding their own proposals. Self-motivation drives the most productive researchers in both applied and basic research; burn-outs are present in both.

Applied researchers have the satisfaction of knowing that their research has a concrete benefit for humanity. Basic researchers know that their research may have highly leveraged downstream applications, and that knowledge is an intrinsically worthwhile aspect of culture. What is the value of culture?

> "To assess basic research by its application value would be awful. It would be like assessing the value of the works of Mozart by the sum they bring in each year to the Salzburg Festival." [Lorenz, 1962]

* * *

Every scientist, basic or applied, has an implicit contract with society. Most scientists are paid by either industry or (perhaps indirectly) by state or federal government in the expectation that we will provide rewarding results. Technology is one such result; another is teaching that is enhanced by active participation in science. Basic researchers are in a unique position to recognize ways that their research might be of practical value. For a basic researcher to take salary and support services from the public, while neglecting possible usefulness of that research to society, is fraudulent.

The synergy between academic research and the local economy has not been quantified, but clues can be found in a detailed survey of the economic relationship between Stanford University and Silicon Valley technology. Most notable was the direct personnel influence: one third of the 3000 small companies in Silicon Valley were created by people who were or had been associated with Stanford. Direct technology transfer, though important, was much more modest: only 5% of the technology employed by these companies came directly from Stanford research [Lubkin et al., 1995].

* * *

Changing Goals for Applied and Basic Research

Attitudes toward applied and basic research are not just a concern for individuals; they also affect national policy. When resources are tight, for example, how can a nation set priorities for funding of basic and applied science? How can funding agencies choose among such diverse research areas as subatomic particles and the human genome? One approach is to define the goals of science, from a national perspective [Gomory, 1993]. Setting goals is a powerful basis for decision-making. Unfortunately, the choice of goals for basic and applied research is hotly debated.

The goal of basic research is reliable knowledge of nature, and the goal of applied science is useful knowledge of nature. These objectives are, perhaps, too sweeping to guide science funding. Until recently, U.S. research funding has been guided by the rationale laid out by Vannevar Bush [1945] half a century ago: both basic and applied research inevitably serve the mission of strengthening national security, mainly by promoting national defense but also by increasing self-sufficiency and standard of living. Bush's vision catalyzed the subsequent growth of U.S. research funding and the breadth of supported disciplines. Priorities have gradually shifted toward greater emphasis on health and medicine, but the framework has remained intact until the last decade.

Some recent attempts to redefine U.S. scientific goals [Gomory, 1993; COSEPUP, 1993] appear to me to be based on the following flawed assumption: a nation or company does not need to make the discoveries; it just needs to be poised to use the discoveries of others. Gomory [1993] and numerous government officials extend this idea even farther, arguing that the purpose of science is industrial competitiveness. If so, perhaps basic science can be reduced to a support service for applied science. A minority [e.g., Jarrard, 1994; Cohen and Noll, 1994] respond that it would be a mistake to redefine the goal of science as industrial competitiveness.

Industrial competitiveness is essential to the economic welfare of the U.S., it is a high national priority, and it is a modern mantra. It is not -- and has never been -- the primary objective of scientific research. Making industrial competitiveness the purpose of applied research defines resulting industrial improvements in other countries as liabilities, not assets. Both individual companies and individual countries benefit from total technological growth, even without competitive advantage.

Pragmatism, not naïveté, suggests the following criterion for national science funding: return on investment, not relative advantage. How much money should a nation invest in basic and applied science? As with all potential investments, the first step is to evaluate return on investment:

> "Science is an endless and sustainable resource with extraordinary dividends." [Executive Office of the President, 1994]

> "Basic research . . . has been an astounding success, whether measured in terms of understanding natural phenomena or improving material wealth and living standards of the world." [Gomory, 1993]

> Many economic studies have investigated the relationship of R&D to productivity, and "the main conclusions from their work are that more than half the historical growth in per capita income in the U.S. is attributable to advances in technology and that the total economic return on investment in R&D is several times as high as that for other forms of investment." [Cohen and Noll, 1994]

With confidence in return on investment, one then invests as much as one can afford. More funded research will lead inevitably to more discoveries, increased productivity, and a higher standard of living.

* * *

Resolution: Bridging the Gap

The changing national priorities for basic and applied research affect research in many ways. The long-term cost-effectiveness of research remains unchallenged. The current focus of concern is, instead, on maximizing the efficacy and speed with which basic-research findings are transferred to the marketplace. One resulting trend is a reallocation of resources, with a higher proportion going to applied research. Today about half of the Ph.D.'s in science and engineering are employed outside the academic environment -- a substantial increase since the 1970's [National Science Foundation, 1994]. Another response is simply a more conscientious linkage between basic research and its potential applications to quality of life (e.g., in industry, professions, and health).

Research funding is changing. The proportion of projects funded entirely by a single grant from a federal agency is dropping. Increasingly, funding agencies are requiring cost sharing and collaboration with private industry. Joint projects between academic researchers and businesses are sprouting at an unprecedented rate, as both groups discover that carefully framed collaborative projects permit individuals to maintain their own objectives and benefit from broader expertise. For example, companies are recognizing the R&D leverage inherent in using faculty expertise and faculty-generated government cost sharing.

Universities are implementing mechanisms for assuring technology transfer and cooperative research among faculty, students, and local business. Some examples are student internships, graduate-student summer jobs in local industry, undergraduate research opportunity programs, university research parks, technology transfer offices, and seed money for research oriented toward technology development.

> "To feed applied science by starving basic science is like economising on the foundations of a building so that it may be built higher." [Porter, 1986]

How far will the pendulum of transformation in research funding swing? The rift between applied and basic research is decreasing, but is there still too much emphasis on basic research? At state and national levels, some are asking whether we really need and can afford the research universities.

Both research and graduate-level teaching make the same major demand on an individual's time: to be up-to-the-minute in a specialized and rapidly growing field. Whereas textbooks are fine for the undergraduate level where well-established 'facts' are taught, graduate-level teaching and research must be at the cutting edge where new ideas are being proposed, evaluated, and rejected. Active researchers are the best guides in this frontier, where the graduate student must learn to travel.

Graduate study is an apprenticeship. Like undergraduate education, it includes some texts and lectures. Unlike undergraduate education and trade schools, most graduate and professional programs require an individually tailored interplay of supervised yet independent study, a learning-by-doing that develops both specialized knowledge and a combination of competencies and work attitudes. Effective graduate-level apprenticeship requires a mentor, research facilities, identification of feasible and decisive research topics, and usually research funding. The research component of a research university is designed to provide exactly these requirements.

These two aspects of graduate study, apprenticeship and evaluation of new ideas, make graduate study less amenable to distance learning and electronic teaching than is undergraduate study. The combination of personal attention and electronic technology is, in contrast, at the heart of graduate education in a research university.

* * *

Big Science versus Little Science

As the geometric growth in number of scientists collides with the linear growth in available science funding, debate is inevitable about where the scarce resources should go. Much of this debate has centered on the issue of big versus little science. More accurately, since there is a continuum between the two, the question is: what are the optimum proportions between large and small projects within a discipline, in order to maximize the scientific payoff per dollar expended?

Big science can be in the form of a single multi-investigator project or research thrust, or a large facility that is used by many researchers for their individual small-science projects. Multibillion dollar examples of the former are the Human Genome Project, (cancelled) supercollider, and space station, though even within these projects there are many moderate-scale subprojects. Examples of large facilities for small-scale projects are telescopes, oceanographic ships, supercomputers, and Antarctic research stations.

Proponents of small science point out that most major discoveries have been a product of small research groups working with modest funding. Such projects are very cost-effective, because most of the money goes to scientists rather than to the equipment and technicians that generally consume most big-science dollars. Advocates of large science accept these arguments, but they claim that the waves of small science have merely washed around some key problems that were too expensive to tackle. Now these problems are the most critical issues remaining to be solved; they can no longer be bypassed.

Most scientists do small science. If science were democratic, many of the big science projects could not fly. Thus the proponents of the largest projects seek a different constituency; they also solicit line-item funding that does not obviously reduce small-science funding. Successful proponents of big science tend to be well known senior scientists who already head large groups and who are on committees charged with outlining new directions for a field. Younger and less famous scientists feel left out.

> This week my closest colleague at Columbia University won the largest grant that Columbia had ever received. Yet most of my friends there are less successful 'soft-money' researchers, who doubt that they will be able to write enough successful proposals to provide their own salaries next year. Also this week, cosmologists are ecstatic over the results of the big-science COBE satellite: the big bang theory has received remarkably strong confirmation, through a mapping of the original subtle heterogeneity of its radiation. Is there a more fundamental scientific question that the origin of the universe, the mother of all singularities?

> Debate over the Human Genome Project was often personal. James Wyngaarden, who was head of the National Institutes for Health when NIH started funding of the project, said "Most knowledgeable people and most eminent scientists are solidly behind [the project]. The ones who are critical are journeymen biochemists who may be having a hard time competing themselves." James Watson, Nobel laureate and previous head of the program at NIH, said, "It's essentially immoral not to get it done as fast as possible." [Angier, 1992]

Polarization and alienation are hazards of the battle. The big-science projects generate another hazard: grand expectations. Virtually all funded proposals make confident predictions of valuable results; optimism and a modest amount of oversell are almost prerequisites for funding. Most of these projects will be somewhat fruitful, partly because the investigators are free to react and refocus their research to avoid obstacles and exploit discoveries, but most projects will also deliver less than the proposals promised. Small-science projects can get away with this because there is safety in

numbers: a few projects will be stunningly rewarding, and the combination of these breakthroughs and many smaller successes creates rapid small-science progress.

Big science, however, lacks this safety in numbers. If a single big-science project fails, public reaction to the 'wasted taxpayers' money' can hurt all scientists' reputations and funding prospects. Such was the initial impact of the Hubble space telescope. Fortunately, NASA corrected its deficiencies, and recent Hubble results have been breathtaking.

* * *

Ego and the Scientific Pecking Order

> "Go, wondrous creature! mount where Science guides,
> Go, measure earth, weigh air, and state the tides;
> Instruct the planets in what orbs to run,
> Correct old Time, and regulate the Sun.
> . . . Go, teach Eternal Wisdom how to rule --
> Then drop into thyself, and be a fool!" [Pope, 1733]

The scientific pecking order is another manifestation of the attitude of "me, in competition with them; me, better than them; me, rather than them." Like chickens, some scientists seem to be obsessed with climbing an imagined pecking order. Those 'below' such a scientist see a scornful user of their efforts; those 'above' such a scientist see a productive team player.

Beyond the local interpersonal pecking order is a broader pecking order of professions that is remarkably fluid in its ability to place one's personal field at the apex. One common pecking order of scientific superiority is *'hard' sciences (i.e., physical sciences)* **>** *social sciences*. Within the physical sciences *physics* **<>***mathematics* (of course depending on whether one is a physicist or mathematician), and *physics & math* **>** *astronomy* **>>** *other physical sciences*. For example, the following provocative 'joke' by Rutherford [Blackett, 1962] makes a non-physicist's blood boil: "All science is either physics or stamp collecting." *Academics* **>** *applied scientists* of industry, because of the hasty generalization that the latter are materialists first and scientists only second. *Applied scientists* **>** *academics*, because of the hasty generalization that the latter are marginally useful ivory-tower dabblers. For example, the applied scientist Werner von Braun said [Weber, 1973], "Basic research is what I'm doing when I don't know what I am doing."

Theoreticians **>** *experimentalists*, because the latter are less intelligent grunt workers. *Experimentalists* **>** *theoreticians*, because the latter are out of touch with reality and think that 'data are confirmed by the model.' Oliver [1991], for example, claims that theories are useless except as an "organization of observations" and that "observation is the ultimate truth of science." *Full-time researchers* ('full-time scientists') **>** *college teachers* ('part-time scientists') **>** *high school teachers*, though in reality the latter may have the most highly leveraged impact on science and the least glory. *Professor* **>** *assistant professor* **>** *lecturer* **>** *student*, because seniority is more important than originality. *Ph.D. researchers* **>** *technicians* > *scientific administrators*, because the latter are not 'true scientists', though they may be just as essential for science.

All of these hierarchies are counterproductive and hypocritical. They are counterproductive because the pecking instinct allows only one at the top of each of the many hierarchies; we all must be both pecked and peckers. This defensive ego building is successful in creating a feeling of superiority only by careful editing of perceptions to focus downward. It is also counterproductive because time and energy are wasted worrying about where one is. The scientific pecking order is hypocriti-

cal because it is an *ex post facto* justification. Almost no one picks their scientific specialty based on the above considerations (a possible exception is the choice between applied and basic research). Fortunately, we pick a field instead because it fascinates us most, and we pick a job within that field because it somehow suits us most. We might almost say that the scientific field chose us, and we obeyed in spite of rational reasons to the contrary.

For the explorers of nature, there are no box seats, no upper-balcony seats. Remember Dedekind's postulate: every segment of a numeric series, however small, is itself infinite. Similarly, within every scientific field are infinities to be explored.

> "And there is no trade or employment but the young man following it may become a hero,
>
> And there is no object so soft but it makes a hub for the wheeled universe." [Whitman, 1892]

How much of our pecking is, like chickens, a reaction to being pecked? How much is ego massage? How much is our need to have a status commensurate with our years of effort? How much is the desire to give a rational explanation for an emotionally inspired career choice?

The scientist's banes are egoism and egotism. The scientific pecking order is one manifestation of egoism, the self-centered practice of valuing everything only in proportion to one's own interest. Egotism, pride, and self-conceit are enhanced by a combination of peer recognition, the value placed by society on intelligence and technology, and one's own false sense of the importance of their contributions to science.

Egotism is not in proportion to peer recognition. Peer recognition and fame can aggravate egotism, but it need not do so. For example, on receiving the 1925 Gold Medal from the Royal Astronomical Society of London, Albert Einstein's response was:

> "He who finds a thought that lets us penetrate even a little deeper into the eternal mystery of nature has been granted great grace. He who, in addition, experiences the recognition, sympathy, and help of the best minds of his time, has been given almost more happiness than a man can bear." [Einstein, 1879-1955]

Too often, "he who finds a thought that lets us penetrate even a little deeper into the eternal mystery of nature" thinks that he is hot shit. Perhaps this is the egotistical trap: we fool ourselves into thinking that we are wresting the secrets away from Nature or God and therefore we must be godlike. Campanella, a 17th century Italian philosopher, described man as:

> "a second god, the first God's own miracle, for man commands the depths, mounts to heaven without wings, counts its moving bodies and measures their nature. . . . He knows the nature of the stars. . . and determines their laws, like a god. He has given to paper the art of speech, and to brass he has given a tongue to tell time." [cited by Smith, 1930]

How much of the pride and ego of modern science is a cultural phenomenon? For example, the boasting about mental powers and control stems partly from the Renaissance feeling that humans are master of the earth. Contrast the ancient Greek perspective that wonder is more appropriate than self-conceit, because people can never achieve the ideals revealed by science. Empedocles [5th century B.C.] said:

> "And having seen [only] a small portion of life in their experience, they soar and fly off like smoke, swift to their dooms, each one convinced of only that very thing

> which he has chanced to meet, as they are driven in all directions. But each boasts of having seen the whole."

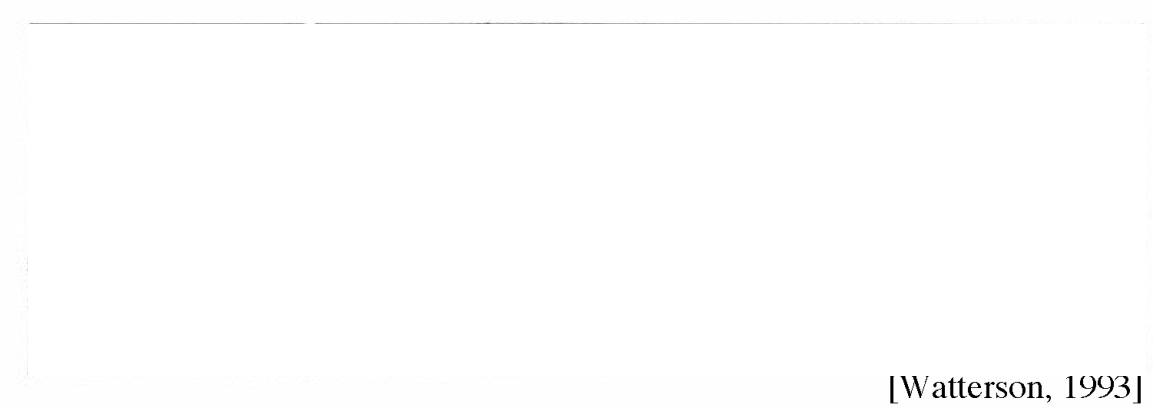
If scientists allow themselves to be seen as Prometheus giving the power of fire to humanity, then they may start thinking of themselves as demigods. J. Campbell [1988b] reminds us that "Man did not weave the web of life, he is merely a strand in it." Bronowski [1973] speaks feelingly and eloquently of the danger of this scientific egotism:

> "[Mathematician] Johnny von Neumann was in love with the aristocracy of intellect. And that is a belief which can only destroy the civilisation that we know. If we are anything, we must be a democracy of the intellect. We must not perish by the distance between people and government, between people and power, by which Babylon and Egypt and Rome failed. And that distance can only be conflated, can only be closed, if knowledge sits in the homes and heads of people with no ambition to control others, and not up in the isolated seats of power."

The luster of an individual's contributions to science is tarnished by the near certainty that some other scientist would have made the same contribution sooner or later. One can help erect the scaffolding of the scientific cathedral, but the scaffolding later will be torn down and forgotten. One can cling to the comfortable fantasy of scientific immortality, but today's scientific breakthrough will be tomorrow's naïveté.

> "Voltaire, when complemented by someone on the work he had done for posterity, replied, 'Yes, I have planted four thousand trees'. . . Nearly a score of centuries ago, Marcus Aurelius reminded us that, 'Short-lived are both the praiser and the praised, the rememberer and the remembered.'" [Teale, 1959]

> "I returned, and saw under the sun, that the race is not to the swift, nor the battle to the strong, neither yet bread to the wise, nor yet riches to men of understanding, nor yet favour to men of skill; but time and chance happeneth to them all." [Solomon, ~1000 B.C., Ecclesiastes 10:11]



[Watterson, 1993]

Chapter 10: The Scientist

Let's conclude by turning our gaze inward. Knowing that science thrives on a diversity of styles and techniques, can we nevertheless identify dominant patterns of behavior, ethics, and motivations?

> "One thing I have learned in a long life: that all our science, measured against reality, is primitive and childlike -- and yet it is the most precious thing we have." [Einstein, 1879-1955]

> Isaac Newton [1642-1727], a man known more for his arrogance than for humility, said near the close of his life:
>
> "I do not know what I may appear to the world; but to myself I seem to have been only like a boy playing on the sea-shore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me."

Scientists' Characteristics

The traits possessed by successful scientists are seldom examined systematically in college or graduate school. They are not the traits that one would choose in an idealized characterization of a scientist. Nor are they revealed by courses and tests. Most courses and tests emphasize temporary fact accumulation, a worthy but largely unnecessary acquisition in an age of ready access to reference information. Some personal characteristics are so pervasive among scientists that they appear to be essential for scientific success. Others are common, advantageous, but not essential.

Essential Characteristics

• **persistence**: This necessary characteristic encompasses traits such as dogged perseverance, patience, tenacity, thoroughness and singleness of purpose. Perhaps, attainment of a Ph.D. demonstrates persistence more than any other capability. As a musician friend told me, daily practice encounters peaks of surging progress and bogs of apparent stagnation. Both are transitory stages that must be outlasted; persistence is the bridge. For scientific success, persistence must continue beyond research and through publication.

> "Nothing in the world can take the place of persistence. Talent will not; nothing is more common than unsuccessful men with talent. Genius will not; unrewarded genius is almost a proverb. Education alone will not; the world is full of educated derelicts. Persistence and determination alone are omnipotent." [Hoenig, 1980]

> "Let me tell you the secret that has led me to the goal. My only strength resides in my tenacity." [Pasteur, 1822-1895,a]

Persistence is not always a virtue. One needs to recognize when to let go -- unlike the weasel, reduced to a skull, found with jaws still imbedded in the throat ruff of a living eagle. It is naive optimism to think that any problem can be solved just by working harder. If a problem is reaching diminishing returns, it should be abandoned. Perhaps the problem is posed wrongly:

> Matthiessen [1978] says that the Buddha "cried out in pity for a yogin by the river who had wasted twenty years of his human existence in learning how to walk on water, when the ferryman might have taken him across for a small coin."

Persistence in a technically difficult experiment is commendable; persistence in investigating a discredited hypothesis is not. If advocacy of an opinion has become counterproductive, then adapt. Nevertheless, far more scientists have failed because of insufficient persistence than because of excessive persistence.

> ". . . let us run with patience the race that is set before us." [Hebrews 12:1]

• **curiosity**: The desire to know more, an inquisitiveness that is not satisfied with surface explanations, is the ratchet of scientific progress.

> Jonas Salk [1990] said that he spent his life "reading the scriptures of nature. . . I began to tease out the logic of the magic that I was so impressed by."

The scientist's curiosity is not passive; it is an active embrace of nature:

> "I come down to the water to cool my eyes. But everywhere I look I see fire; that which isn't flint is tinder, and the whole world sparks and flames." [Dillard, 1974]

• **self-motivation:** Internal drive to work is a product of job enjoyment. Self-motivation is scarce in most types of jobs [Terkel, 1974], frequent in professions, and nearly universal among productive scientists. Single-minded drive undoubtedly increases effort, but self-motivation seems to have more impact than effort can account for. Self-motivated scientists, who may do only part-time research because of teaching or administrative responsibilities, can produce more than full-time scientists who have lost their internal drive (e.g., because management does not value their work).

Self-motivation can be overdone: I and many scientists whom I know are stress junkies, who are stimulated so much by 'emergencies' that they seem to create such situations even when a rapid pace is unnecessary. To a stress junkie, efficiency and productivity are additional sources of job satisfaction.

> Volcanologist Maurice Krafft, who was later killed by Unzen Volcano, said " I would say that if one truly specializes in explosive volcanoes then it's not worth contributing towards retirement, and that if one makes it to retirement it's a little suspicious. It means that he really didn't do his job conscientiously." [Williams and Montaigne, 2001]

Productivity has become a cliché of the business world, but productivity is not just a national or industrial goal. It is a personal skill. Computer expertise and efficient fact finding are tangible forms of individual scientific productivity; more essential and less tangible aspects are problem-solving ability, quantitative reasoning, and self-motivation. Quantity of publications is the most commonly used measure of productivity [Maddox, 1993]. Its virtues are simplicity and objectivity, but scientific impact does not depend on number of publications.

> "Every man, every civilization, has gone forward because of its engagement with what it has set itself to do. The personal commitment of a man to his skill, the intellectual commitment and the emotional commitment working together as one, has made the Ascent of Man." [Bronowski, 1973]

• **focus:** Focus is the ability to spot the crux among a morass of detail and then stay concentrated on it, without being distracted or sidetracked. Focus assures that the target receives all of the attention needed. Lack of focus is evidenced by tendencies toward incompleteness, inefficiency, overlooked significant details, grasshopper science, and panic reaction to setbacks.

> Thanks to physicist Richard Feynman [1985], I now associate focus with chocolate pudding. During a period of his life when he went out to dinner frequently, a waiter asked him what he would like for dessert. Suddenly he considered how much of his life was wasted in thinking about that trivial question, so he decided that henceforth the answer would always be chocolate pudding! Focus does not tolerate needless distractions.

• **balance between skepticism and receptivity:** A critical attitude is essential; all data and interpretations must be evaluated rather than simply accepted. Yet it is equally essential to achieve a balance between skepticism and receptivity: willingness to propose speculative hypotheses that may be proved wrong, tempered by ability to weed out the incorrect hypotheses. One must be receptive to novel concepts or results, rather than greeting the new with a 'fight-or-flight' reaction of dismissive criticism. The critical filter that rejects everything as insufficiently proved robs science both of joy and of raw materials for progress.

This balance is manifest also by a blend of optimism and realism. Optimism and enthusiasm for new ideas are contagious and powerful, if accompanied not by a casual confidence that effort alone will find a solution, but by a problem-solving mentality and preparation for potential obstacles.

Common Characteristics

Many prospective scientists think that love of science and high intelligence are the two primary characteristics needed to permit them to be successful scientists. This idealized picture of science can lead to disillusionment or worse. Love of science and high intelligence are neither necessary nor sufficient, though they are the springboard of most scientific careers.

• **fascination with the beauty of nature:** We may not use words such as 'beauty of nature'; we may try (at least outwardly) to maintain the myth of objectivity. Yet we revel in the elegance and wonder of the natural world, and we choose occupations that further our opportunities for appreciation of it.

> "I am among those who think that science has great beauty. . . A scientist in his laboratory is not only a technician but also a child placed in front of natural phenomena which impresses him like a fairy tale." [Marie Curie, 1937]

> Konrad Lorenz [1962] described the prerequisites to success in the field of animal behavior as follows:
>
> "To really understand animals and their behavior you must have an esthetic appreciation of an animal's beauty. This endows you with the patience to look at them long enough to see something. Without that joy in just looking, not even a yogi would have the patience. But combined with this purely esthetic characteristic, you must have an analytical mind."

• **love of science:** Love of science is a greater spur to productivity than any manager can offer. Love of science, love of discovery, and enthusiasm for science are contagious; they are nurtured by scientific interactions. Most scientists are inclined to be somewhat forgiving of weaknesses in those colleagues who passionately love science.

If you were beginning a career again, would you pick the same type of work? The answer to this question was 'yes' for 86-91% of physical scientists, 82-83% of lawyers and journalists, 41-52% of those in skilled trades (printers, autoworkers, and steelworkers), and only 16-21% of those in unskilled trades (assembly-line steelworkers and autoworkers) [Blauner, 1960]. Jobs with the highest levels of worker satisfaction are those that are esteemed by society, that allow both personal control of decisions and unsupervised work, and that involve teams [Blauner, 1960]. Scientific careers provide all of these.

I learned how atypical the scientist's job satisfaction is when I told my half-brother, who was an insurance salesman, that I love my work; he laughed and told me not to bullshit him. Sometimes the exhilaration with science is so overpowering that I break out in a silly grin. Then I remember, consciously or unconsciously, the scientist's distrust of sentimentality. I transform the grin into a knowing smile and dryly remark, "It's a dirty job, but somebody has to do it; don't they?"

• **above-average intelligence**: This characteristic is almost essential, but a scientist with only average intelligence can succeed by excelling in the other traits of scientists. Genius is not required. Among those with an IQ > 120, IQ shows little relation to either scientific innovation or productivity [Simonton, 1988]. Genius without the other needed qualities is insufficient for scientific success.

> Srinivasa Ramanujan was a mathematics genius in 19th-century India. He was rich enough to receive a high school education, a few books, and live as a scholar. Yet for most of his life he was completely cut off from virtually all mathematics literature and knowledge. He worked on his own, and mathematicians are still deciphering and applying his work [Gleick, 1992d]. How much more could he have accomplished as part of the science community? How many geniuses never see a book?

Most of us equate IQ scores with intelligence, but IQ scores predict success in school, not in life. Career and family success is forecast more successfully with tests that model constructive thinking, problem solving, and persuasion, with and without emotional distractions. In contrast, IQ tests evaluate specific types of verbal and mathematical ability. They do not evaluate how well these will be applied to the often ambiguous and open-ended problems of real life, where ability to react to crises and manage one's emotions are just as essential as IQ [Goleman, 1992a].

• **imagination**: Imagination is necessary for insight and even for the everyday problem solving that is intrinsic to most science. Almost all scientists are unusually imaginative, but the unimaginative can produce valuable science in the form of careful hypothesis testing. Individuals who have imagination but lack a critical attitude can be cranks; they cannot be scientists. When imagination is combined with both will and a vision of what is achievable, the result can be formidable: "We choose to go to the moon" [J. F. Kennedy, 1960 speech].

• **desire to improve**: "Boredom could be an important stimulus to evolution among the animals" [Calvin, 1986], because it leads to trials of a variety of different behaviors. Like curiosity, dissatis-

faction with the *status quo* certainly is a stimulus to scientific progress. This dissatisfaction is manifested by boredom, the appeal of the mysterious, and the desire to improve circumstances.

> "The most beautiful experience we can have is the mysterious. It is the fundamental emotion that stands at the cradle of true art and true science." [Einstein, 1879-1955]

The desire to improve encompasses both oneself and one's environment:

> "How thankful I should be to fate, if I could find but one path which, generations after me, might be trodden by fellow members of my species." [Lorenz, 1962]

• **aggressiveness:** Aggressive scientists tend to be highly successful and productive. Science is an obstacle course of puzzles, experimental problems, and bureaucratic hurdles, and success requires an aggressive unwillingness to be stopped by such obstacles. I am cautious, however, about interactions with aggressive scientists, as most of them seem to have trouble finding a balance between ethics and aggressiveness. Ethical barriers are not just problems to be overcome, and other scientists are not just tools to be used for furthering one's progress.

Style determines whether aggressiveness is an asset. For example, we see quite different aggressive styles every day on the highway: some drivers fight the traffic, whereas others go with the flow much of the time, while anticipating congestion and seizing opportunities.

• **self-confidence:** Self-confidence fosters a willingness to face challenges and a constructive optimism, relatively free of worries about the opinions of others and about whether the problem can be solved. Both self-motivation and self-confidence are needed if one is to lead a scientific discipline into new productive directions, rather than just following along with the majority. Self-confidence inspires acceptance of one's opinions by others, in spite of scientists' claims that they are influenced only by the evidence, not by the presentation.

* * *

Scientists are subject to many of the same fears as most people. They fear mediocrity, completing a life of science only to conclude that they had little or no significant impact on science. They fear humiliation, being proved wrong in print or, worse yet, being shown to have made some mistake that 'no real scientist should make.' They fear that someone else will get the credit for their discoveries. They fear that they cannot keep up with the pace of science and are being left behind [Sindermann, 1987].

Perhaps instead they should fear that they have lost proportion: that they are sacrificing too much of their personal life to science, that they have abandoned some ethical values because those values hampered achievement of scientific objectives.

* * *

> "What is then the quality which enables some men to achieve great things in scientific research? For greatest achievements men must have genius -- that elusive quality that so often passes unrecognized, while high ability receives reward and praise. But for achievement genius is not enough, and, for all but the greatest achievements, not necessary. What does appear essential for real achievement in scientific research is a combination of qualities, by no means frequent, but commoner than is genius. It

> seems that these qualities are clarity of mind, a combination of imagination and caution, of receptivity and skepticism, of patience and thoroughness and of ability to finalize, of intellectual honesty, of a love of discovery of new knowledge and understanding, and of singleness of purpose. Of these the most important is the love of discovery of new knowledge and understanding. If any young readers, contemplating scientific research as a profession, do not feel this love . . . scientific research is not for them." [Freedman, 1950]

> "What can I wish to the youth of my country who devote themselves to science? *Firstly, gradualness*. About this most important condition of fruitful scientific work I never can speak without emotion. Gradualness, gradualness and gradualness. Learn the ABC of science before you try to ascend to its summit. Never begin the subsequent without mastering the preceding . . . Do not become the archivists of facts. Try to penetrate the secret of their occurrence, persistently search for the laws which govern them. *Secondly, modesty*. Never think that you already know all. However highly you are appraised, always have the courage to say of yourself – I am ignorant. Do not allow haughtiness to take you in possession. Due to that you will be obstinate where it is necessary to agree, you will refuse useful advice and friendly help, you will lose the standard of objectiveness. *Thirdly, passion*. Remember that science demands from a man all his life. If you had two lives that would be not enough for you. Be passionate in your work and your searchings." [Pavlov, 1936]

* * *

These generalizations concerning characteristics of scientists are subjective, based on my and others' personal observations. In contrast, Rushton [1988] summarizes the results of several objective statistical analyses as follows:

> "Scientists differed from nonscientists in showing high general intellectual curiosity at an early age and in being low in sociability. . . Eminent researchers [were] . . . more dominant, self sufficient, and motivated toward intellectual success. . . In summary, the impression that emerges of the successful research scientist is that of a person less sociable than average, serious, intelligent, aggressive, dominant, achievement oriented, and independent."

* * *

Cooperation or Competition?

Both cooperation and competition are integral aspects of scientific interaction. Joint projects combine diverse, specialized expertise to promote research success. For many scientists, competition provides a motivation to excel. This drive to win is particularly effective for those researchers who can pace themselves, putting out a burst of extra effort on those occasions when it can make the decisive difference between being a discover and being a confirmer of others' discoveries.

The choice between scientific cooperation and competition is a daily one, involving conscious or unconscious decisions on style of interactions with scientific peers. Most scientists simplify this decision-making by adopting a strategy that provides the decision. Perhaps the strategy is to cooperate with all other scientists; perhaps it is to compete with everyone over everything. More likely, the individual always cooperates with a few selected scientists and competes with others. Whatever viable strategy is selected, we should recognize its consequences.

The survival value of successful competition is almost an axiom of evolutionary theory. Why, then, has cooperation survived evolutionary pressure, in humans as well as in many other species? Kinship theory is the usual explanation. According to kinship theory, a genetically influenced strategy such as cooperation is evolutionarily viable if it helps a substantial portion of one's gene pool to survive and reproduce, even if the cooperator dies. A classic example is the sterile worker honeybee, which commits suicide by stinging. Altruism of parents for offspring is easy to explain, but kinship theory also successfully predicts that altruism would be high among all members of an immediate family and present throughout an inbred tribe. Sacrifice for an unrelated tribe member may improve future treatment of one's children by tribe members.

Modified kinship theory can account for many manifestations of cooperation and competition among scientists. An *us/them* perspective can be developed among members of a company, university, or research group. Thus a member of a National Science Foundation proposal-review panel must leave the room whenever a proposal from their home institution is under discussion. Here the health or reputation of an institution is an analogue for genetic survival. Similarly, a clique of scientists with the same opinion on a scientific issue may cooperate to help defeat a competing theory.

For scientists facing the decision of cooperation or competition with a fellow scientist, kinship theory is not a particularly useful guide. A more helpful perspective is provided by the concept of an evolutionarily stable cooperation/competition strategy. Evolution of a cooperation/competition strategy, like other genetic and behavioral evolutions, is successful only if it fulfills three conditions [Axelrod and Hamilton, 1981]:

• **initial viability**. The strategy must be able to begin by gaining an initial foothold against established strategies.

• **robustness**. Once established, the strategy must be able to survive repeated encounters with many other types of strategy.

• **stability**. Once established, the strategy must be able to resist encroachment by any new strategy.

Axelrod and Hamilton [1981] evaluated these three criteria for many potential cooperative/competitive strategies by means of the simple game of Prisoner's Dilemma [Rapoport and Chammah, 1965]. At each play of the game, two players simultaneously choose whether to cooperate or defect. Both players' payoffs depend on comparison of their responses:

My choice	Other's choice	My score	Explanation
cooperate	defect	0	**S**ucker's disadvantage
defect	defect	1	**N**o-win mutual defection
cooperate	cooperate	3	**R**eward for mutual cooperation
defect	cooperate	5	**C**ompetitive advantage

When the game ends after a certain number of plays (e.g., 200), one wants to have a higher score than the opponent. But even more crucial if the game is to be an analogue for real-life competition and cooperation, one seeks the highest average score of round-robin games among many individuals with potentially varied strategies.

The optimum strategy in Prisoner's Dilemma depends on both the score assignments and the number of plays against each opponent. The conclusions below hold as long as:

• $S < N < R < C$, i.e., my defection pays more than cooperation on any one encounter, and cooperation by the opponent pays more to me than his or her defection does;

• R > (C+S)/2, i.e., cooperation by both pays more than alternating exploitation; and

• I neither gain nor lose from my opponent's scoring (e.g., if I were to gain even partially from his gains, then continuous cooperation would be favored).

If one expects to play only a single round against a specific opponent, then the optimum strategy in Prisoner's Dilemma is to *always defect*. Similarly, in a population of individuals with no repeat encounters or within a species incapable of recognizing that an encounter is a repeat encounter, constant competition is favored over cooperation. More relevant to interactions among scientists, however, is the case of many repeat encounters where one remembers previous encounters with a given 'opponent'. It is this situation that Axelrod and Hamilton [1981] modeled by a computer round robin tournament, first among 14 entries and then among 62 entries of algorithm strategies submitted by a variety of people of different professions. Subsequent computer studies by various investigators simulated the process of biological evolution more closely, incorporating variables such as natural selection (higher birth rate among more successful strategies) and mutation.

In nearly all simulations, the winner was one of the simplest of strategies: *tit for tat*. *Tit for tat* cooperates on the first move, then on all subsequent moves duplicates the opponent's preceding move. Axelrod and Hamilton [1981] call *tit for tat* "a strategy of cooperation based on reciprocity." When *tit for tat* encounters a strategy of *all defect*, it gets burned on its first cooperative move but thereafter becomes a strategy of *all defect*, the only viable response to an *all defecter*. *Tit for tat* does much better against itself than *all defect* does against itself, and *tit for tat* also does much better against various other strategies, because mutual cooperation pays off more than mutual defection.

Axelrod and Hamilton [1981] prove that *tit for tat* meets the success criteria of initial viability, robustness, and stability for Prisoner's Dilemma, and they argue that *tit for tat* is also a successful evolutionary strategy in various species from human to microbe (its reactive element does not require a brain). Some of their examples are highly speculative, while others such as territoriality ring true. Individuals in adjacent territories develop stable boundaries ('cooperation'), but any attempt by one individual to encroach is met by aggression by the other. In contrast to this dominantly *tit for tat* behavior with the same individual, one-time encounters with encroaching strangers are consistently met by aggression (*all defect*).

Tit for tat does have two weaknesses. First, a single accidental defection between two *tit for tat* players initiates an endless, destructive sequence of mutual defections. Second, a *tit for tat* population can be invaded temporarily by persistent cooperators. An alternative strategy – *win-stay, lose-shift* – copes with these situations more successfully [Nowak and Sigmund, 1993]. This strategy repeats its former move if it was rewarded by a high score (opponent's cooperation); otherwise, it changes its move. The strength of this strategy stems from the fact that cooperation by the opponent is more beneficial than their defection. *Win-stay, lose-shift* quickly corrects mistakes, and it exploits chronic cooperators.

It's incredible that we scientists make decisions – sometimes difficult, sometimes emotion-laden – based on strategies similar to those used by some single-celled organisms. Success of *tit for tat* and *win-stay, lose-shift* in computer games of Prisoner's Dilemma does not imply that these strategies are appropriate guides for interactions with fellow scientists. Experience shows that the extremes of total cooperation and total competition are also viable for some scientists, although the 'hawks' do take advantage of the 'doves'. Some doves react to being repeatedly taken advantage of by becoming either bitter or hawkish. *Tit for tat* seems like a more mature reaction to being exploited than does rejection of all cooperation.

Which strategy is best for science? Both cooperation and competition are stimulating to scientific productivity, and in different individuals either appears to be able to give job satisfaction by fulfilling personal needs. Communication of scientific ideas is clearly a win-win, or non-zero-sum

game [Wright, 2000]. On the other hand, academic science is being forced into an overly competitive mode by the increasing emphasis on publication records for both funding and promotion decisions [Maddox, 1993]. Personally, I enjoy cooperation more and I unconsciously seem to use *tit for tat*, achieving cooperation most of the time without the sucker's disadvantage.

> "And though I have the gift of prophecy, and understand all mysteries, and all knowledge; and though I have all faith, so that I could remove mountains, and have not charity, I am nothing." [1 Corinthians 13]

* * *

Science Ethics

Personal and professional ethics are not distinguishable; all ethics are personal. A scientist must make ethical decisions with care, not only because they affect self image but also because, as Sindermann [1987] has pointed out, scientific reputations are fragile.

Some rules of scientific ethics are universal, and others are subjective. All require personal judgment. Not all of the ethical opinions that follow can claim consensus. Another perspective, and one which has been subjected to much wider review, is given in the excellent pamphlet, "On Being a Scientist" [Committee on the Conduct of Science, 1989]. Scientists are not democratic; most insist on deciding personally whether a rule warrants following, rather than accepting the majority vote. Imagine yourself, for example, in the following situations; what would your decision be in each case?

Research project:

• You have just completed a study on the effect of X on Y. Nineteen of the twenty data points exhibit a very close relationship between X and Y, but something seems to be wrong with one data point: it is far different from the pattern. Should you omit it from your publication entirely, include it and explain that you consider it to be anomalous, or include it just like the other data?

• In your publication you cite relevant studies by others. Should you devote just as much discussion to studies that are inconsistent with your conclusions as to studies that are consistent?

• You have reached an insight inspired by reading a preprint, a pre-publication copy of a scientific article. Should you immediately publish the idea, giving credit to the preprint?

• For the paper that you are writing, should you include as authors people who have made useful suggestions? People who did only 5% of the work? People who did substantial work but disagree with your analysis or conclusions?

• Your graduate student has selected, carried out, and written up a project. You provided funding and guidance. What should the authorship be?

Research-related issues:

• Is it OK to make a personal copy of software, if you cannot afford to purchase it? Is it OK to buy one copy of a program, then install it on all of the computers in your lab?

• You are filling out a travel expense form. It forbids claiming an item (e.g. tips) that you consider to be a legitimate expense, or you failed to get a receipt for some item and you are not allowed re-

imbursement without a receipt. Should you increase some other category by an equivalent amount? Should you claim full per diem when your actual daily expenses were substantially less?

• You are writing the budget for a proposal. Knowing that the funding agency routinely cuts budgets by 10-30%, should you pad the proposal budget by 20%?

• A funding agency has announced that it seeks proposals on some subject. You are doing work on a quite similar subject. Should you submit a proposal, recasting your work in terms of the desired research? If funded, is it OK to continue in the original research area rather than focusing entirely on the area desired by the funding agency?

• In submitting a proposal, you know that including one subproject will substantially increase the chances of proposal funding but that the subproject is not really viable. Should you include it anyway? Should you say that you will accomplish more than you realistically expect to achieve?

Applied vs. basic research:

• Is it selfish or even ethical to carry out a government-funded basic research career, if you think that your research has absolutely no practical value?

• If your research has potential practical applications that you approve of, should you suggest that your employer get a patent or should you start an independent company that gets a patent?

• If your research has potential practical applications that you disapprove of, should you reveal them?

Every ethical decision must be weighed personally and subjectively. Before making a final decision on any ethical issue, it is worthwhile to consider the issue from the standpoint of Kohlberg's [1981, 1984] criterion for mature moral judgment: *does the judgment hold regardless of which position one occupies in the conflict?* It may be worth reviewing your decisions on the ethical questions above, this time pretending that you were a person affected by the decision rather than the one making the decision. Kohlberg's criterion sounds almost like a generalization of "Do unto others as you would have them do unto you." The habit of applying Kohlberg's criterion is analogous to the habit (or skill) of objectively evaluating the effect of data on various hypotheses, without regard for which hypothesis one favors [Kuhn et al., 1988].

* * *

Verbal, if not always behavioral, unanimity prevails on three ethical issues: fraud, intellectual honesty and theft of ideas.

Fraud and falsification of data are so inimical to scientific method that almost never do scientists succumb to their lure of quick rewards. Even a single case of scientific fraud, when publicized, does unimaginable damage to the credibility of scientists in general, for the public cannot confirm our findings; they must trust them. Fraud also slows the advance of a scientific field, for experiments seldom are exactly replicated, and fraud is not suspected until all alternative explanations have been eliminated.

> "The scientific mind is usually helpless against a trained trickster. Because a man has mastered the intricacies of chemistry or physics is no qualification for him to deduce how the Chinese linking rings, for instance, seem to melt into each other, passing metal through solid metal. Only a knowledge of the rings themselves can reveal the secret. The scientific mind is trained to search for truth which is hidden in the mys-

teries of nature -- not by the ingenuity of another human brain." [Houdini, 1874-1926]

Intellectual honesty must be a goal of every scientist. As we saw in Chapter 6, people tend to ignore evidence that diverges from expectations. We must fight this tendency; continued awareness and evaluation of possible personal biases is the best weapon. Intellectual honesty requires that we remain alert to conflicts of interest, whenever we review proposals and manuscripts, and wherever objectivity and personal advancement clash. Intellectual honesty requires that we face weaknesses as well as strengths of data, hypotheses, and interpretations, without regard for their origin, invested effort, or potential impact on our beliefs.

"Thou shalt not steal," and the currency of scientists is not money or objects but ideas. **Intellectual plagiarism**, the attempt to take credit for the ideas of others, is clearly unacceptable, but its boundaries are indefinite. Most scientists feel that:

• It is not OK to initiate a research project inspired by communications from another scientist in a letter, conversation, lab visit, or preprint, unless the other scientist has specifically encouraged you to do so. Ask permission, but weigh the other's response to decide whether they really favor your jumping in or they simply feel obliged to say yes.

• It is OK to initiate a research project inspired by another scientist in a scientific talk for which abstracts are published. One should refrain from publishing a manuscript on this project, however, until the other scientist has published.

• It is not OK to let your research plans be affected in any way by either a proposal or a manuscript that you have been sent for review.

• It is OK to jump into a research area as soon as it has been published. The authors have no right to keep the field to themselves, and they do have a head start.

• When mentioning a previously published idea in a publication, reference the originator unless the idea has become common knowledge.

Intellectual plagiarism is more often suspected than deliberately practiced. Ideas frequently stem from interactions with others. In such cases, the combination of two perspectives deserves credit for development of the idea, not the person who first verbalizes it. Perhaps the idea is not even verbalized during the discussion, yet one of the individuals later 'realizes' the idea when solitarily thinking about the subject. Menard [1986], reviewing the formative days of the geological paradigm of plate tectonics, found that simultaneous 'independent' discoveries were remarkably common.

* * *

Publication

"To study, to finish, to publish." [Benjamin Franklin, 1706-1790]

Communication of results, particularly via publication, is an essential part of a scientist's life. I could describe the highly ritualized design of most modern publications: introduction, experimental techniques, observations, and conclusions. I am more intrigued, however, by the contrast between publications, which are dry and rational, and publication experiences, which can be heavily emotion-laden. Let us examine briefly the publication experiences of some of our greatest scientific forebears: Euclid, da Vinci, Newton, Darwin, Mendel, and Einstein.

* * *

Pythagoras founded Greek mathematics and especially geometry in about 550 B.C.. He and his Pythagorean school made the geometry one of the greatest accomplishments of Greek science. For systematically expounding and expanding Pythagorean geometry, however, we owe thanks to Euclid. In Alexandria in about 300 B.C., he wrote Elements of Geometry, and until recent years it was the most translated and copied book in history except for the Bible [Bronowski, 1973]. Many famous scientists in these two millennia thanked Euclid's book for showing them the beauty of what Pythagoras called the 'language of nature'.

* * *

Leonardo da Vinci (1452-1519) combined the eye of an artist with the curiosity and analytic ability of a scientist. He epitomized the breadth and depth of the Italian Renaissance, by the scope of subjects and the novelty of perspectives in his notes. He was self-taught, with no intellectual training and therefore minimal limiting dogma [Goldstein, 1988].

Unfortunately, Leonardo made absolutely no contribution to contemporary scientific knowledge. He did not interact with scientists and he did not publish anything. His now-famous notes were private, written backwards to prevent casual reading by others. If a researcher publishes nothing and thereby makes no contribution whatsoever to the field of science, can that person even be called a scientist? Such questions are as fruitless as the question of whether a scientist-administrator is a scientist. Certainly Leonardo was an inspiring example to later scientists. Certainly Leonardo's lack of scientific communications to his peers was a heartbreaking loss to science.

* * *

The greatest scientific book of all time is Principia Mathematica, completed by Isaac Newton in 1687. Newton's paradigm of the physics of motion united terrestrial and planetary motions with simple mathematical laws. He elegantly demonstrated the ability of theoretical physics to derive precise predictions of empirically observable phenomena. Yet Newton was so insecure and so incapable of dealing with the criticisms of others that he nearly failed to make his findings public. He completed much of the work of Principia many years before publishing it. Only Edwin Hubble's constant encouragement and partial financing eventually compelled Newton to produce Principia.

Twenty years earlier, when Newton began his work on gravitation, he developed the calculus. Rather than publish calculus, he kept it secret, using it to make several discoveries but then couching the presentation of these discoveries in ordinary mathematics. In about 1676 Gottfried Leibniz developed calculus independently. Newton, convinced that Leibniz had somehow stolen the idea from him, started a bitter feud.

Although Newton was undoubtedly one of the most brilliant scientific minds in history, his insecurity fostered arrogance and prevented him from distinguishing scientific criticism from personal criticism. He was ridiculed, and he responded by trying to discredit and destroy other scientists. Personal weakness damped, at least temporarily, his scientific impact. Fortunately, he did publish.

* * *

Alfred Russel Wallace and Charles Darwin independently discovered the theory and mechanism of evolution. Both recognized the phenomenon of evolutionary divergence, based on extensive observations as a naturalist (particularly in South America). Both spent years seeking a mechanism for this divergence, and both credited their discovery of that mechanism to reading Malthus.

Wallace called the mechanism survival of the fittest and Darwin called it natural selection. But Darwin's insight was in 1838 and Wallace's was in 1858.

Darwin, like Newton, was reluctant to publish but even more reluctant to see someone else get the credit for 'his' discovery. Fortunately, other scientists arranged for Wallace and Darwin to present their results in talks at the same meeting. Darwin's Origin of Species, published in 1859, stunned the scientific world with the weight of its diverse data. Its conclusions were so radical that overwhelmingly compelling data were essential. Darwin left the task of arguing the case to others.

Wallace is largely forgotten today, but he had experienced far greater disappointment than seeing Darwin receive much of the credit for the theory of evolution: after spending four years collecting animal specimens in the Amazon, he lost everything when the ship home caught fire.

> "With what pleasure had I looked upon every rare and curious insect I had added to my collection! How many times, when almost overcome by the ague, had I crawled into the forest and been rewarded by some unknown and beautiful species! How many places, which no European foot but my own had trodden, would have been recalled to my memory by rare birds and insects they had furnished to my collection!
>
> "And now everything was gone, and I had not one specimen to illustrate the unknown lands I had trod or to call back the recollection of the wild scenes I had beheld! But such regrets I knew were vain, and I tried to think as little as possible about what might have been and to occupy myself with the state of things which actually existed." [Wallace, 1853].

To Wallace, 1858 brought two joys: he solved a problem that had obsessed him for years, and he was personally responsible for the public awareness of the revolutionary concept of evolution. The most important thing that he brought back from South America was in his mind, not in flammable boxes.

* * *

Gregor Mendel undertook and published one experiment in his life. He used measurements of characteristics of sweet peas to lay out the basic pattern of genetic inheritance. The results overthrew the conventional theory that offspring inherit traits intermediate between their two parents; they demonstrated instead that offspring inherit each trait from only one parent, in predictable integer proportions.

"Mendel published his results in 1866 in the Journal of the Brno Natural History Society, and achieved instant oblivion. No one cared. No one understood his work" [Bronowski, 1973]. He picked an obscure journal, he failed to distribute copies of his paper to biologists, he was a monk rather than a professional scientist because he had flunked out of the university, and his research was before its time. Thirty years passed before biologists were ready to appreciate Mendel's paper.

* * *

Paradigm change can be explosively rapid on the time scale of evolution of a scientific field, yet ploddingly slow on the time scale of an individual scientist. While working full time at the patent office in 1905, Albert Einstein published five revolutionary papers: light quantized like particles rather than waves, diffusion-based estimates of the size of molecules and of Avogadro's number, Brownian motion (a final confirmation of the existence of atoms), the special theory of relativity, and conversion of mass into energy.

He later submitted the diffusion paper to the University of Zurich as a potential doctoral thesis. It was rejected as too short; Einstein added one sentence and resubmitted it, and it was accepted. But

when in 1907 he wanted to leave the patent office and become a university lecturer, he needed to have an approved inaugural thesis. He submitted his 1905 paper on special relativity to Bern University; the paper was rejected as 'incomprehensible' [Hoffmann, 1972].

* * *

In considering the publication examples above, there are lessons that I accept intellectually but do not fully practice. I have not always published my research results. Nor have I always sold my results successfully. These examples increase my motivation to complete projects by publishing effectively. By revealing familiar personality traits, these examples increase my sense of community with past and present scientists.

* * *

A Scientist's Life: Changing Motivations

Career motivations, within science or other professions, are not static. They evolve – sometimes radically. For a significant proportion of scientists, parts of the composite scientific life below may be familiar.

She chose science, or was chosen by it, while she was a child. Through childhood and undergraduate years, her fascination with science was a love of learning how the world works. Books, including textbooks, were the road. 'Facts' were collected uncritically and enthusiastically.

> Naturalist Edwin Way Teale was six years old when he first experienced, in a patch of forest, the fascination with nature that guided his life. In later years he tried unsuccessfully to refind that patch of forest, but the power of the initial experience remained with him:
>
> "For me, the Lost Woods became a starting point and a symbol. It was a symbol of all the veiled and fascinating secrets of the out-of-doors. It was the starting point of my absorption in the world of Nature. The image of that somber woods returned a thousand times in memory. It aroused in my mind an interest in the ways and the mysteries of the wild world that a lifetime is not too long to satisfy." [Teale, 1959]

In graduate school, her motivations changed:

> "Enough of Science and of Art;
> Close up those barren leaves;
> Come forth, and bring with you a heart
> That watches and receives."
> [Wordsworth, 1798]

After two decades of experiencing science through textbooks, she found a compelling alternative to texts: personal scientific discovery. Second-hand knowledge paled by comparison. The interpretations of others were subjective and required personal evaluation, mainly on scientific grounds. Observation and insight were an intoxicating combination. Competing and being first were part of the game.

In her thirties and forties, being first almost became the game. Recognition brought responsibilities that were essential to science. Time was short: it was more efficient to advance science through administration, management, and the training of students. Students took over the time-consuming data collection, but her scientific planning, data interpretation, hypothesis generation and

insight continued unabated. Recognition and power brought their own rewards. The opinions of others, like dependent variables, could be modified to achieve her objectives.

Love of science seems to be universal among new scientists. Yet it fades in some scientists, particularly those who become managers and administrators. Perhaps individuals move from research into administration partly because of waning thrill of scientific research. Perhaps they move first, as reluctant draftees who are called on to serve a need, and they find later that love of science is being supplanted by fresher job satisfactions such as recognition and power. Failures we all can afford. The cost of success, for many, is loss of wonder.

> "If I would be a young man again and had to decide how to make my living, I would not try to become a scientist or scholar or teacher. I would rather choose to be a plumber or a peddler in the hope to find that modest degree of independence still available under present circumstance." [Einstein, 1954]

Few of these motivational changes were based on systematic strategic planning of her career. More often she simply reacted to the many victories, frustrations, and emotional fireworks of day-to-day life. Yet she perceived the true importance of these when suddenly she faced her own mortality.

> "Sometimes one finds in fossil stones the imprint of a leaf, long since disintegrated, whose outlines remind us how detailed, vibrant, and alive are the things of this earth that perish." [Ackerman, 1990]

During the Cuban missile crisis, we faced the world's mortality. After the crisis, we reassured each other, saying "I'm glad that's over." We returned to our old lives but found that we had somehow changed. Facing mortality changes one ineffably: the critical becomes trivial, and new priorities emerge. In the blazing light of awareness of death, the inessential and peripheral are burned away. Few things remain: love and living science are two.

Between now and my death is an opportunity. How shall I use it?

> After the albatross was killed, and before it was avenged:
> "The fair breeze blew, the white foam flew,
> The furrow followed free;
> We were the first that ever burst
> Into that silent sea."
> [Coleridge, 1798]

* * *

Process and Product

They say that Tantalus was punished by the gods, doomed to see a branch of fruit tree waving in the wind just beyond his reach, doomed to see the waters retreat from him each time he dipped his palm to drink, and thus consigned to be forever hungry and thirsty. Millennia later, the Buddha sat beneath a bo tree, determined to remain there until he gained knowledge. Both were tantalized by their objective; only one eventually embraced the path itself, learning the archer's skill of knowing when to pull and when to let go. Today, eager to quench our appetites, we scientists grasp for the same fruit of knowledge.

> "I dreamed that I floated at will in the great Ether, and I saw this world floating also not far off, but diminished to the size of an apple. Then an angel took it in his hand and brought it to me and said, 'This must thou eat.' And I ate the world." [Emerson, 1840]

"Knowledge is our destiny," said Bronowski [1973], and sometimes I am similarly goal-oriented in expressing my motivation toward science: I accumulate facts in hopes of finding understanding; I accumulate understandings in hopes of finding wisdom. Certainly these are aspects of my drive for living science, but perhaps the ends are merely a justification for the means. I think that Joseph Campbell [1988a] perceived a deeper obsession in his parable of the 'motivation' of the grass in a lawn:

> The grass grows, and yet every week or so a human comes along and ruthlessly mows it, annihilating all of the week's progress. Does the grass think, "Oh, for Pete's sake, I give up!" Of course not. For the mower, as for the mown, it goes on, toward ends unknown.
>
> "It bothers some people that no matter how passionately they may delve, the universe remains inscrutable. 'For my part,' Robert Louis Stevenson once wrote, 'I travel not to go anywhere, but to go. . . The great affair is to move.' . . . It began in mystery, and it will end in mystery, but what a savage and beautiful country lies in between." [Ackerman, 1990]

[Watterson, 1993]

Those who are living science love the *process* of science -- the unique synergy of control and freedom, of skepticism and innovation. They love to use all of the scientific methods and try to dodge their pitfalls. Only rarely does the lightning flash of insight course through them, more often they feel the satisfaction of a successfully diagnostic experiment, and daily they overcome minor hurdles.

> At times, when I lived in Alaska, the brightness of the night sky kept me awake. Last night, its darkness did the same. How can the night sky be dark? If the universe is infinite, then shouldn't it be uniformly bright, lit by an infinite number of stars in every direction? This 'dark-sky paradox' has puzzled astronomers for more than a century, and it has been 'solved' more than a dozen times [Gleick, 1992a]. The modern solution begins by reminding us that what we see in the night sky is the photons currently reaching our eyes, recording events that happened on different stars light-years ago. And how far back in time can we see? No farther than the 12 billion-year-ago big-bang origin of the universe. Stars more than 12 billion light-years away are invisible to us today, because the light hasn't reached us yet. It goes on, toward ends unknown.

I'm no astronomer, but still I wonder. Didn't Einstein show that two stars cannot move apart faster than the speed of light? How, then, can the unseen stars be so far away? Who can relax and sleep, if the universe is breaking its speed limit? Is the universe infinite, and are those unseen stars there in the black portions of sky? Walt Whitman [1892], as usual, had the keenest vision: "The bright suns I see and the dark suns I cannot see are in their place."

References

Abel, J. J., 1930, The education of the superior student, *J. Chemical Education, 7*, pp. 283-293.

Abelard, P., 1122, *Sic et Non*, extracts in: J. H. Robinson (transl.), 1904, *Readings in European History*, Ginn & Co.: Boston, pp. 450-451.

Achinstein, P., 1985, The method of hypothesis: What is it supposed to do, and can it do it? *In* P. Achinstein and O. Hannaway (Eds.), *Observation, Experiment and Hypothesis in Modern Physical Science,* M.I.T. Press: Cambridge, MA, pp. 127-145.

Ackerman, D., 1990, *A Natural History of the Senses*, Vintage Books: New York.

Alhazen, ~1000 A.D., *The Optics of Ibn Al-Haytham*, 1989, Warburg Institute, Univ. of London: London.

Angier, N., 1992, The codebreakers, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 169-174.

Archimedes, ~287-212 B.C., *In* T. L. Heath (transl.), 1952, *The Works of Archimedes Including the Method*, Encyclopedia Britannica: Chicago.

Aristotle, 384-322 B.C., *Metaphysica*, *In* W. D. Ross (Ed.), 1908-1952, *The Works of Aristotle (12 vols.)*, Clarendon Press: Oxford.

Arnaudin, M. W., J. J. Mintzes, C. S. Dunn, and T. H. Shafer, 1984, Concept mapping in college science teaching, *Journal of College Science Teaching, 14*, pp. 117-121.

Arnauld, A., 1662, *Logic, or, The Art of Thinking ("The Port-Royal Logic")*, J. Dickoff and P. James (transl.), 1964, Bobbs-Merrill: Indianapolis.

Augustine, St., 354-430, a, *In* H. Paolucci (Ed.), J. F. Shaw (transl.), 1961, *The Enchiridion on Faith, Hope, and Love*, Regnery Gateway: Chicago.

Augustine, St., 354-430, b, cited by: R. Chambliss, 1954, *Social Thought from Hammurabi to Comte*, Henry Holt & Co.: New York.

Ausubel, D. P., J. D. Novak, and H. Hanesian, 1978, *Educational Psychology: A Cognitive View*, Holt, Rinehart, & Winston: New York.

Axelrod, R., and W. D. Hamilton, 1981, The evolution of cooperation, *Science, 211,* pp. 1390-1396.

Bacon, F., 1561-1626, *Advancement in Learning*, *In* 1952, *Great Books of the Western World, v. 30*, Encyclopedia Britannica, Inc.: Chicago.

Bacon, F., 1620, *Novum Organum*, R. Ellis and J. Spedding (transl.), 1900?, George Routledge & Sons, Ltd., London.

Bacon, R., ~1270, *In* T. L. Davis (transl.), 1923, *Roger Bacon's Letter Concerning the Marvelous Power of Art and of Nature and Concerning the Nullity of Magic*, Chemical Publ. Co.: Easton, PA.

Bailey, N. T. J., 1967, *The Mathematical Approach to Biology and Medicine*, Wiley: New York.

Baker, A., 1970, *Modern Physics and Antiphysics*, Addison-Wesley: Reading, MA.

Bauer, H. H., 1994, *Scientific Literacy and the Myth of the Scientific Method*, Univ. of Illinois Press: Urbana, IL.

Bernard of Chartres, ~1150, cited by: E. R. Fairweather, 1956, *A Scholastic Miscellany: Anselm to Ockham*, Westminster Press: Philadelphia.

Bernard, C., 1865, *An Introduction to the Study of Experimental Medicine*, H. C. Green (transl.), 1925, Dover Publ.: New York.

Bernstein, J., 1982, *Science Observed; Essays Out of My Mind*, Basic Books: New York.

Beveridge, W. I. B., 1955, *The Art of Scientific Investigation*, Vintage Books: New York.

Blackett, P. M. S., 1962, Memories of Rutherford. *In* J. B. Birks (Ed.), *Rutherford at Manchester*, Heywood & Co.: London, pp. 102-113.

Blake, W., ~1803, *Auguries of Innocence*, *In* A. Ostriker (Ed.), 1977, *William Blake, The Complete Poems*, Penguin Books: Harmondsworth, UK.

Blatty, W. P., 1972, *The Exorcist*, Bantam Books: Toronto.

Blauner, R., 1960, Work satisfaction and industrial trends in modern society, *In* W. Galenson and S. M. Lipset (Eds.), *Labor and Trade Unionism: an Interdisciplinary Reader*, John Wiley and Sons: New York, pp. 339-360.

Biondi, A. M., 1980, About the small cage habit, *J. Creative Behavior*, *14*, pp. 75-76.

Boghosian, B. M., 1990, Computational physics on the connnection machine, *Computers in Physics, 4(1),* pp. 14-32.

Bonnichsen, R., and A. L. Schneider, 1995, Ancient hair, and the DNA embedded in it, might reveal when and how the Americas were settled – but not if some Native Americans can help it, *The Sciences*, May/June, pp. 27-31.

Boyd, R. N., 1985, Observations, explanatory power, and simplicity: Toward a non-Humean account, *In* P. Achinstein and O. Hannaway (Eds.), *Observation, Experiment and Hypothesis in Modern Physical Science*, M.I.T. Press: Cambridge, pp. 47-94.

Brehaut, E., 1912, *An Encyclopedist of the Dark Ages: Isidore of Seville*, Columbia Univ. Press: New York.

Bronowski, J., 1973, *The Ascent of Man*, Little, Brown & Co.: Boston.

Brown, W. L., Jr., 1987, Punctuated equilibrium excused: the original examples fail to support it, *Biol. J. of the Linnean Society*, *31*, pp. 383-404.

Browne, M. W., 1992, A star is dying, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 46-49.

Bruner, J. S., and L. Postman, 1949, On the perception of incongruity: a paradigm, *J. Personality, 18*, pp. 206-223.

Buck, P. S., 1962, *A Bridge for Passing*, John Day Co.: New York.

Burns, E. M., 1963, *Western Civilizations, Their History and Their Culture (6th ed.)*, W. W. Norton: New York.

Bush, V., 1945, *Science -- the Endless Frontier*, report to the President.

Calvin, W. H., 1986, *The River That Flows Uphill*, Sierra Club Books: San Francisco.

Campbell, J., with B. Moyers, 1988a, *Joseph Campbell and The Power of Myth* (videorecording), Mystic Fire Video: New York.

Campbell, J., with B. Moyers, 1988b, *The Power of Myth*, Doubleday: New York.

Cannon, W. B., 1945, *The Way of an Investigator*, W. W. Norton: New York.

Carlyle, T., 1795-1881, cited by: A. L. Mackay, 1991, *A Dictionary of Scientific Quotations*, Inst. of Physics Publ.: Bristol, UK.

Carnap, R., 1966, *Philosophical Foundations of Physics; An Introduction to the Philosophy of Science*, Basic Books, Inc.: New York.

Chambliss, R., 1954, *Social Thought from Hammurabi to Comte*, Henry Holt & Co.: New York.

Cheng, P., and K. Holyoak, 1985, Pragmatic reasoning schemas, *Cognitive Psychology, 17*, pp. 391-416.

Clausewitz, C. von, 1830, *On War (Vol. 3)*, J. J. Grahan (transl.), 1956, Routledge & Kegan Paul Ltd.: London.

Cohen, L. R., and R. G. Noll, 1994, Privatizing public research, *Sci. American, 271(3),* pp. 72-77.

Coleridge, S. T., 1798, The Rime of the Ancient Mariner, *In* C. M. Coffin (Ed.), 1954, *The Major Poets: English and American*, Harcourt, Brace & World: New York, pp. 274-292.

Committee on the Conduct of Science, National Academy of Sciences, 1989, *On Being a Scientist*, National Academy Press: Washington, DC.

Conant, J. B., 1947, *On Understanding Science; An Historical Approach*, Yale Univ. Press: New Haven.

COSEPUP (Committee on Science, Engineering, and Public Policy), 1993, *Science, Technology, and the Federal Government: National Goals for a New Era*, National Academy Press: Washington, DC.

Cowell, A., 1992, The threat of the new, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 241-244.

Crombie, A. C., 1953, *Robert Grosseteste and the Origins of Experimental Science, 1100-1700*, Clarendon Press: Oxford.

Cunningham, J., 1988, Techniques of pyramid-building in Egypt, *Nature*, *332(6159)*, pp. 22-23.

Curie, E., 1937, *Madame Curie*, Doubleday, Doran & Co., Inc.: Garden City, NY.

Dante Alighieri, 1313-1321, *In* C. E. Norton (transl.), 1941, *The Divine Comedy of Dante Alighieri*, Houghton Mifflin Co.: Boston.

Darwin, C., 1876, *Autobiography*, *In* F. Darwin (Ed.), 1897, *The Life and Letters of Charles Darwin, Vol. I*, D. Appleton & Co.: New York.

Davy, H., 1840, *The Collected Works of Sir Humphrey Davy*, J. Davy (Ed.), Smith, Elder & Co.: London.

Derry, G. N., 1999, *What Science is and How it Works*, Princeton Univ. Press: Princeton, NJ.

Descartes, R., 1629, *Rules for the Direction of the Mind*, E. S. Haldane and G. R. T. Ross (transl.), 1952, Encyclopedia Britannica: Chicago.

Descartes, R., 1637, *Discourse on Method, and Meditations*, L. J. Lafleur (transl.), 1960, Liberal Arts Press: New York.

Diaconis, P., and F. Mostelle, 1989, Methods for studying coincidences, *J. Am. Stat. Assoc., 84*, pp. 853-861.

Dillard, A., 1974, *Pilgrim at Tinker Creek*, Harper's Magazine Press: New York.

Dixon, W. J., and F. J. Massey Jr., 1969, *Introduction to Statistical Analysis (3rd ed.)*, McGraw-Hill: New York.

Doyle, A. C., 1893a, *Memoirs of Sherlock Holmes*, A. L. Burt: New York.

Doyle, A. C., 1893b, *A Study in Scarlet*, Ward, Locke & Bowden: New York.

Doyle, A. C., 1917, *His Last Bow: Some Reminiscences of Sherlock Holmes*, Penguin: Harmondsworth, UK.

Einstein, A., 1879-1955, cited by: B. Hoffmann, 1972, *Albert Einstein, Creator and Rebel*, Viking Press: New York.

Einstein, A., 1954, Correspondence to the editor, *The Reporter, 11(9)*, p. 8.

Eiseley, L., 1978, *The Star Thrower*, New York Times Books Co.: New York.

Eldredge, N., and S. J. Gould, 1972, Punctuated equilibria: an alternative to phyletic gradualism, *In* T. J. M. Schopf (Ed.), *Models in Paleobiology*, Freeman, Cooper: San Francisco.

Emerson, R. W., 1840, *In* A. W. Plumstead and H. Hayford (Eds.), 1960, *The Journals and Miscellaneous Notebooks of Ralph Waldo Emerson*, *v. 7*, Belknap Press: Cambridge, MA.

Empedocles, 5th century B.C., *The Poem of Empedocles*, B. Inwood (transl.), 1992, Univ. of Toronto Press: Toronto.

Ennis, R., 1969, *Logic in Teaching*, Prentice-Hall, Inc.: Englewood Cliffs, NJ.

Executive Office of the President, Office of Science and Technology Policy, 1994, *Science in the National Interest*, U.S. Govt. Printing Office: Washington, DC.

Fairweather, E. R., 1956, *A Scholastic Miscellany: Anselm to Ockham*, Westminster Press: Philadelphia.

Feynman, R. P., 1985, *Surely You're Joking, Mr. Feynman! Adventures of a Curious Character*, Bantam Books: Toronto.

Feynman, R. P., 1988, *What Do You Care What Other People Think? Further Adventures of a Curious Character*, W. W. Norton & Co.: New York.

Fisher, R. A., and F. Yates, 1963, *Statistical Tables for Biological, Agricultural and Medical Research (6th ed.)*, Hafner Publ. Co.: New York.

Fowler, W. S., 1962, *The Development of Scientific Method*, Pergamon Press: Oxford.

Franklin, B., 1706-1790, cited by: A. L. Mackay, 1991, *A Dictionary of Scientific Quotations*, Inst. of Physics Publ.: Bristol, UK.

Freedman, P., 1950, *The Principles of Scientific Research*, Public Affairs Press: Washington, D.C.

Gaither, C. C., and A. E. Cavazos-Gaither, 2000, *Scientifically Speaking: A Dictionary of Quotations*, Institute of Physics Publishing: Bristol.

Galileo Galilei, 1564-1642, *In* Redondi, P., 1987, *Galileo: Heretic*, Princeton Univ. Press: Princeton NJ.

Galison, P., 1985, Bubble chambers and the experimental workplace, *In* P. Achinstein & O. Hannaway (Eds.), *Observation, Experiment and Hypothesis in Modern Physical Science*, M.I.T. Press: Cambridge, MA, pp. 309-373.

Gerard, R. H., 1946, The biological basis of imagination, *Sci. Monthly, 62*, pp. 477-499.

Gibbon, E., 1787, *The Decline and Fall of the Roman Empire*, W. Whitman (Ed.), 1943, Wise: New York.

Giere, R. N., 1983, Testing theoretical hypotheses, *In* J. Earman (Ed.), *Testing Scientific Theories, Minnesota Studies in the Philosophy of Science, 10*, pp. 269-298.

Gilbert, G. K., 1886, The inculcation of scientific method by example, with an illustration drawn from the Quaternary geology of Utah, *American J. Science (3rd series), 31*, pp. 284-289.

Gleick, J., 1987, *Chaos; Making a New Science*, Penguin Books: New York.

Gleick, J., 1992a, Why is the night sky dark? *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 29-32.

Gleick, J., 1992b, Chicken little, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 50-53.

Gleick, J., 1992c, Faking it, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 293-297.

Gleick, J., 1992d, The enigma of genius, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 357-362.

Goldstein, T., 1988, *Dawn of Modern Science*, Houghton Mifflin: Boston.

Goleman, D., 1992a, IQ isn't everything, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 140-145.

Goleman, D., 1992b, The making of memory, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 135-139.

Goleman, D., 1992c, The tenacity of prejudice, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 126-130.

Gomory, R. E., 1993, Goals for the federal role in science and technology, *Physics Today, 46(5)*, pp. 42-45.

Gould, S. J., 1981, *The Mismeasure of Man*, W. W. Norton: New York.

Gould, S. J., 1990, Enigmas of the small shellies, *Natural History, 1990(10)*, pp. 6-17.

Gregory, R. L., 1966, *Eye and Brain; The Psychology of Seeing*, McGraw-Hill: New York.

Harris, S., 1970, *What's So Funny about Science?* William Kaufmann, Inc.: Los Altos, CA.

Harris, S., 1982, *What's So Funny about Computers?* William Kaufmann, Inc.: Los Altos, CA.

Hartmann, W. K., 1983, *Moons and Planets (2nd ed.)*, Wadsworth Publ.: Belmont, CA.

Haskins, C. H., 1927, *The Renaissance of the Twelfth Century*, World-New American Library: New York.

Hastorf, A. H., and H. Cantril, 1954, They saw a game: a case study, *Journal Abn. Soc. Psych., 49,* pp. 129-134.

Heisenberg, W., 1927, Principle of indeterminacy, *Z. Physik*, *43*, pp. 172-198.

Heisenberg, W., 1958, *Physics and Philosophy; The Revolution in Modern Science*, Harper: New York.

Helmholtz, H. von, 1891, cited by: Koenigsberger, L., 1906, *Hermann von Helmholtz*, F. A. Welby (transl.), Clarendon Press: Oxford.

Helmholtz, H. von, 1903, *Vorträge u. Reden, 5th Aufl.*, F. Vieweg und Sohn: Braunschweig.

Hemingway, E., 1940, The Snows of Kilimanjaro, *In* A. Gingrich (Ed.), *The Bedside Esquire*, Grossett & Dunlap: New York, pp. 253-274.

Heminway, J., 1983, *No Man's Land, A Personal Journey into Africa*, Warner Books: New York.

Heraclitus, ~550-475 B.C., cited by: G. S. Kirk, 1970, *Heraclitus; The Cosmic Fragments*, Cambridge Univ. Press: London.

Hesse, H., 1923, *Siddharta*, H. Rosner (transl.), 1951, New Directions Publishing Co.: New York.

Hilts, P. J., 1992, Heterosexuals and AIDS, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 190-194.

Hoenig, S. A., 1980, *How to Build and Use Electronic Devices Without Frustration, Panic, Mountains of Money, or an Engineering Degree (2nd ed.)*, Little, Brown: Boston.

Hoffman, M. S. (Ed.), 1990, *The World Almanac and Book of Facts*, Pharos Books: New York.

Hoffmann, B., 1972, *Albert Einstein, Creator and Rebel*, Viking Press: New York.

Hofstadter, D. R., 1985, *Metamagical Themas: Questing for the Essence of Mind and Pattern*, Basic Books: New York.

Holton. G. J., and D. R. D. Roller, 1958, *Foundations of Modern Physical Science*, Addison-Wesley Pub.: Reading, MA.

Hoover, K. R., 1988, *The Elements of Social Scientific Thinking (4th ed.)*, St. Martin's Press: New York.

Houdini, H., 1874-1926, cited by: N. L. Browning, 1970, *The Psychic World of Peter Hurkos*, Doubleday: Garden City, NY.

Hubbard, R., 1957, Have only men evolved? *In* Hubbard, R., M. S. Henefin, and B. Fried (Eds.), *Women Look at Biology Looking at Women*, Schenkman Publishing Co.: Cambridge, pp. 7-36.

Hume, D., 1935, *An Enquiry Concerning Human Understanding*, Open Court Publ. Co.: Chicago.

Hurley, P. J., 1985, *A Concise Introduction to Logic*, Wadsworth Publ. Co.: Belmont, CA.

Ibrahim, Y. M., 1991, Arab bus kills Israeli and driver is slain, *New York Times*, Jan. 5, p. L-3.

Ittelson, W. H., and F. P. Kilpatrick, 1951, Experiments in perception, *Sci. Amer., 185(2)*, pp. 50-55.

Jarrard, R. D., 1994, Gomory's 'goals' too narrowly national, *Physics Today, 47(7)*, pp. 13-14.

Jason, G., 1989, *The Logic of Scientific Discovery*, Peter Lang: New York.

Jefferson Starship, 1970, *Have You Seen the Stars Tonite*, *In Blows Against the Empire*, RCA Records: New York.

Jeffrey, R. C., 1985, Probability and the art of judgment, *In* P. Achinstein & O. Hannaway (Eds.), *Observation, Experiment, and Hypothesis in Modern Physical Science*, M.I.T. Press: Cambridge, MA, pp. 95-126.

Jones, L., 1990, *In* Earthquake, *Nova (PBS TV)*.

Kennedy, J. F., 1963, Address to the National Academy of Sciences.

Killeffer, D. H., 1969, *How Did You Think of That? An Introduction to the Scientific Method*, Doubleday: Garden City, NY.

Kohlberg, L., 1981, *Essays on Moral Development: Vol. 1. The Philosophy of Moral Development*, Harper & Row: New York.

Kohlberg, L., 1984, *Essays on Moral Development: Vol. 2. The Psychology of Moral Development*, Harper & Row: New York.

Kolata, G., 1992a, Is alcoholism inherited? *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 229-232.

Kolata, G., 1992b, The burden of proof, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 223-228.

Kropotkin, P. A., 1899, *Memoirs of a Revolutionist*, Houghton, Mifflin & Co.: Boston.

Kuhn, D., E. Amsel, and M. O'Loughlin, 1988, *The Development of Scientific Thinking Skills*, Academic Press: San Diego.

Kuhn, T., 1970, *The Structure of Scientific Revolutions (2nd ed.)*, Univ. of Chicago Press: Chicago.

Kuhn, T., 1977, *The Essential Tension: Selected Studies in Scientific Tradition and Change*, Univ. of Chicago Press: Chicago.

Lakatos, I., 1970, Falsification and the methodology of scientific research programs, *In* I. Lakatos and A. Musgrave (Eds.), *Criticism and the Growth of Knowledge*, Cambridge Univ. Press: Cambridge, U.K., pp. 91-195.

Lambert, K., and G. Brittan, Jr., 1970, *An Introduction to the Philosophy of Science*, Prentice Hall: Englewood Cliffs, NJ.

Langmuir, I., 1928, Atomic hydrogen as an aid to industrial research, *Ind. & Engin. Chem., 20*, pp. 332-336.

Larson, G., 1980, *The Far Side Gallery*, Andrews, McMeel & Parker: Kansas City.

Larson, G., 1985, *Valley of the Far Side*, Andrews and McMeel: Kansas City.

Larson, G., 1987, *The Far Side Observer*, Andrews and McMeel: Kansas City.

Larson, G., 1989, *Wildlife Preserves: A Far Side Collection*, Andrews and McMeel: Kansas City.

Lee, D., 1950, Codifications of reality: lineal and nonlineal, *Psychosomatic Medicine, 12(2)*, pp. 89-97.

Lenzen, V. F., 1938, Procedures of empirical science, *In* O. Neurath, R. Carnap and C. Morris (Eds.), *International Encyclopedia of Unified Science*, University of Chicago Press: Chicago, pp. 279-339.

Leuba, J. H., 1925, *The Psychology of Religious Mysticism*, Harcourt, Brace & Co.: New York.

Loehle, C., 1990, A guide to increased creativity in research – inspiration or perspiration? *BioScience, 40*, pp. 123-129.

Lord, C. G., L. Ross, and M. R. Lepper, 1979, Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence, *J. Personality and Social Psychology, 37*, pp. 2098-2109.

Lorenz, K. Z., 1962, *King Solomon's Ring*, Time Inc.: New York.

Lubkin, G. B., I. Goodwin, R. L. Byer, P. M. Eisenberger, K. Gottfried, L. H. Greene, D. N. Langenberg, E. J. Moniz, D. T. Moore, and S. C. Solomon, 1995, Roundtable: whither now our research universities?, *Physics Today, 48(3)*, pp. 42-51.

Luoma, J. R., 1992, Lake 302, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 245-249.

Maddox, J., 1993, Competition and the death of science, *Nature*, *363*, p. 667.

Magnus, A., ~1250, *De Animalibus*, H. Stadler (Ed.), 1916-1921, Aschendorff: Münster.

Mannoia, V. J., 1980, *What is Science? An Introduction to the Structure and Methodology of Science*, University Press of America: Washington, DC.

Matthiessen, P., 1978, *The Snow Leopard*, Viking Press: New York.

Medawer, P. B., 1967, *The Art of the Soluble*, Methuen & Co.: London.

Medawar, P. B., 1969, Induction and intuition in scientific thought, *Memoirs of the American Philosophical Society*, *75*, pp. 1-62.

Menard, H. W., 1986, *The Ocean of Truth, a Personal History of Global Tectonics*, Princeton Univ. Press: Princeton, NJ.

Mill, J. S., 1930, *System of Logic*, Longmans, Green: New York.

Minkowski, H., 1908, Space and time, *In* H. A. Lorentz, A. Einstein, H. Minkowski, and H. Weyl, 1952, *The Principle of Relativity*, Dover Pub.: London.

Morris, C. W., 1938, Scientific empiricism, *In* O. Neurath, R. Carnap, and C. Morris (Eds.), *International Encyclopedia of Unified Science*, Univ. of Chicago Press: Chicago, pp. 63-75.

Nair, K. R., 1940, Table of confidence interval for the median in samples from any continuous population, *Sankhya, 4*, pp. 551-558.

National Science Foundation, Division of Science Resources Studies, 1994, *Characteristics of Doctoral Scientists and Engineers in the United States*, NSF 94-307.

Neisser, U., 1968, The processes of vision, *Sci. Amer., 219(3)*, pp. 204-214.

Newton, I., 1642-1727, cited by: D. Brewster (Ed.), 1855, *Memoirs of the Life, Writings, and Discoveries of Sir Isaac Newton, Vol. 2*, T. Constable & Co.: Edinburgh.

Newton, I., 1676, cited by: R. K. Merton, 1965, *On the Shoulders of Giants; a Shandean Postscript*, Free Press: New York.

Newton, I., 1687, *Principia*, A. Motte (transl.), 1729, with 1960 revisions by F. Cajori, Univ. Calif. Press: Berkeley.

Newton, I., ~1700, cited by: B. Hoffmann, 1972, *Albert Einstein, Creator and Rebel*, Viking Press: New York.

Nowak, M., and K. Sigmund, 1993, A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game, *Nature*, *364*, pp. 56-58.

Oliver, J. E., 1991, *The Incomplete Guide to the Art of Discovery*, Columbia Univ. Press: New York.

Open University Science Foundation Course Team, 1970, *The Handling of Experimental Data*, The Open University Press: Bletchley, U.K..

Oppenheimer, J. R., 1945, cited by: D. Schroeer, 1972, *Physics and its Fifth Dimension*: *Society*, Addison-Wesley: Reading, MA.

Pasteur, L., 1822-1895, a, cited by: R. J. Dubos, 1950, *Louis Pasteur; Free Lance of Science*, Little, Brown & Co.: Boston.

Pasteur, L., 1822-1895, b, cited by: R. Vallery-Radot, 1927, *The Life of Pasteur*, Garden City Publ. Co.: Garden City, NY.

Pavlov, I. P., 1936, Bequest of Pavlov to the academic youth of his country, *Science, 83*, p. 369.

Penzias, A. A., and R. W. Wilson, 1965, A measurement of excess antenna temperature at 4080 Mc/s, *Astrophys. J., 142*, pp. 419-421.

Percy, W., 1987, *The Thanatos Syndrome*, Ivy Books: New York.

Phelps, E. J., 1899, Jan. 24 speech at Mansion House, London.

Planck, M., 1949, *Scientific Autobiography and Other Papers*, F. Gaynor (transl.), Philosophical Library: New York.

Plato, ~427-347 B.C., a, *In* B. Jowett (transl.), 1937, *The Dialogs of Plato (Republic VIII)*, Random House: New York.

Plato, ~427-347 B.C., b, *In* J. McDowell (transl.), 1973, *Theaetetus*, Clarendon Press: Oxford.

Platt, W., and R. A. Baker, 1931, The relation of the scientific "hunch" to research, *J. Chemical Education, 8*, pp. 1969-2002.

Poincaré, H., 1905, *Science and Hypothesis*, (transl.), 1952, Dover Public.: New York.

Poincaré, H., 1914, *Science and Method*, F. Maitland (transl.), T. Nelson & Sons: London.

Pope, A., 1733, An Essay on Man, *In* C. M. Coffin (Ed.), 1954, *The Major Poets: English and American*, Harcourt, Brace and World: New York, pp. 240-241.

Popper, K., 1959, *Logic of Scientific Discovery*, Basic Books: New York.

Popper, K., 1963, *Conjectures and Refutations; The Growth of Scientific Knowledge*, Basic Books: New York.

Popper, K. R., 1976, The logic of the social sciences, *In* T. W. Adorno, et al., *The Positivist Dispute in German Sociology*, Harper and Row: New York, pp. 87-122.

Porter, G., 1986, Lest the edifice of science crumble, *New Scientist, 111(1524)*, p. 16.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, 1988, *Numerical Recipes in C, The Art of Scientific Computing*, Cambridge Univ. Press: New York.

Rapoport, A., and A. M. Chammah, 1965, *Prisoner's Dilemma; A Study in Conflict and Cooperation*, Univ. of Michigan Press: Ann Arbor.

Reinfeld, F., 1959, *The Complete Chess Course*, Doubleday & Co.: Garden City, NY.

Rifkin, J., with T. Howard, 1980, *Entropy; A New World View*, Viking Press: New York.

Rilke, R. M., 1875-1926, S. Mitchell (transl.), 1984, *Letters to a Young Poet*, Random House: New York.

Rushton, J. P., 1988, Scientific creativity: an individual differences perspective, *J. Social and Biol. Structures, 11*, pp. 140-143.

Russell, B., 1927, *Philosophy*, W. W. Norton & Co.: New York.

Russell, B., 1938, On the importance of logical form, *In* O. Neurath, R. Carnap, and C. Morris (Eds.), *International Encyclopedia of Unified Science*, Univ. of Chicago Press: Chicago, pp. 39-41.

Salk, J., 1990, *Bill Moyer's World of Ideas: Jonas Salk*, PBS TV.

Schwarzschild, B., 2001, Farthest supernova strengthens case for accelerating cosmic expansion, *Physics Today*, *54(6)*, pp. 17-20.

Shah, I., 1970, *Tales of the Dervishes*, E.P. Dutton & Co.: New York.

Shah, I., 1972, *The Exploits of the Incomparable Mulla Nasrudin*, E.P. Dutton & Co.: New York.

Shakespeare, 1600, *As You Like It*, *In* H. Craig (Ed.), 1961, *The Complete Works of Shakespeare*, Scott, Foresman & Co.: Glenview, IL.

Simonton, D. K., 1988, *Scientific Genius*, Cambridge Univ. Press: New York.

Sindermann, C. J., 1987, *Survival Strategies for New Scientists*, Plenum Press: New York.

Smith, P., 1930, *A History of Modern Culture, Vol. I*, Henry Holt & Co.: New York.

Snow, C. P., 1964, *The Two Cultures; and A Second Look*, Mentor Books: New York.

Spencer, H., 1883, *Education: Intellectual, Moral and Physical*, D. Appleton & Co.: New York.

Spencer, J. H., 1983, The largest integer, *In* G. H. Scherr (Ed.), *The Best of the Journal of Irreproducible Results*, Workman Publ: New York, p. 145.

Stevens, W., 1931, Metaphors of a Magnifico, *In Harmonium*, Alfred A. Knopf: New York.

Stevens, W. K., 1992a, Dead heat, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 261-266.

Stevens, W. K., 1992b, Wired, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 282-288.

Sun Tzu, ~500 B.C., *The Art of War*, 1982 transl., Graham Brash (Pte) Ltd: Singapore.

Teale, E. W. (Ed.), 1949, *The Insect World of J. Henri Fabre*, Dodd, Mead: New York.

Teale, E. W., 1959, *Adventures in Nature*, Dodd, Mead: New York.

Terkel, S., 1974, *Working*, Avon: New York.

Thomas, D. H., 1998, *Archaeology (3rd Ed.)*, Harcourt Brace: Fort Worth, TX.

Thomson, J. J., cited by: Thomson, G., 1961, *The Inspiration of Science*, Oxford Univ. Press: London.

Thurstone, L. L., 1925, *The Fundamentals of Statistics*, Macmillan Co.: New York.

Timpane, J., 1991, The poetry of science, *Scientific American, 265(1)*, p. 128.

Toulmin, S., 1967, The evolutionary development of natural science, *Am. Scientist, 55*, pp. 456-471.

Trotter, W., 1941, *The Collected Papers of Wilfred Trotter, F.R.S.*, Oxford Univ. Press: London.

Velikovsky, I., 1967, *Worlds in Collision*, Dell Publ. Co.: New York.

Velikovsky, I., 1977, *Earth in Upheaval*, Pocket Books: New York.

Virgil, 70-19 B.C., *Georgics II*, *In* G. B. Miles, 1980, *Virgil's Georgics; A New Interpretation*, Univ. Calif. Press: Berkeley.

Wallace, A. R., 1853, *Travels on the Amazon and Rio Negro, with an Account of the Native Tribes, and Observations on the Climate, Geology, and Natural History of the Amazon Valley*, Ward: Lock, UK.

Wallas, G., 1926, *The Art of Thought*, Harcourt & Brace: New York.

Wason, P., 1966, Reasoning, *In* B. Foss (Ed.), *New Horizons in Psychology, Vol. 1*, Penguin: Harmondsworth, UK, pp. 135-154.

Watterson, B., 1993, *The Days are Just Packed: A Calvin and Hobbes Collection by Bill Watterson*, Andrews and McMeel: Kansas City.

Weber, R. L., 1973, *A Random Walk in Science*, Institute of Physics: London.

Weyl, H., 1952, *Symmetry*, Princeton Univ. Press: Princeton.

Whitman, W., 1892, Leaves of Grass, *In Walt Whitman; Complete Poetry and Collected Prose*, Literary Classics of United States: New York.

Wilford, J. N., 1992a, Culture shock, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 101-105.

Wilford, J. N., 1992b, Outer limits, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 38-41.

Wilford, J. N., 1992c, Reaching for the stars, *In* R. Flaste (Ed.), *New York Times Book of Science Literacy*, Harper Perennial: New York, pp. 33-37.

William of Conches, ~1150, cited by: M.-D. Chenu (transl. J. Taylor and L. K. Little), 1968, *Nature, Man, and Society in the Twelfth Century*, Univ. of Chicago Press: Chicago.

Williams, S., and F. Montaigne, 2001, *Surviving Galeras*, Houghton Mifflin: Boston.

Wilson, E. B., Jr., 1952, *An Introduction to Scientific Research*, McGraw-Hill: New York.

Wolfe, T., 1936, *The Story of a Novel*, C. Scribner's Sons: New York.

Wordsworth, W., 1798, *In* 1896, *The Complete Poetical Works of William Wordsworth, V. 1*, Macmillan & Co.: London.

Wright, R., 2000, *Non Zero: the Logic of Human Destiny*, Pantheon Books: New York.

Young, H. D., 1962, *Statistical Treatment of Experimental Data*, McGraw-Hill: New York.

Ziman, J. M., 1969, Information, communication, knowledge, *Nature, 224(5217)*, pp. 318-324.

Name Index

Subject Index

E

F

G

H

O

P

R

S

T

V

W