Journal of Modern Applied Statistical Methods
Fall 2002, Vol. 1, No 2, 202-227

Copyright © 2002 JMASM, Inc.
1538 – 9472/02/$30.00

INVITED ARTICLES

Constructive Criticism

Ronald C. Serlin
University of Wisconsin-Madison

Attempts to attain knowledge as certified true belief have failed to circumvent Hume's injunction against induction. Theories must be viewed as unprovable, improbable, and undisprovable. The empirical basis is fallible, and yet the method of conjectures and refutations is untouched by Hume's insights. The implications for statistical methodology is that the requisite severity of testing is achieved through the use of robust procedures, whose assumptions have not been shown to be substantially violated, to test predesignated range null hypotheses. Nonparametric range null hypothesis tests need to be developed to examine whether or not effect sizes or measures of association, as well as distributional assumptions underlying the tests themselves, meet satisficing criteria.

Keywords: Probability, knowledge, satisficing, statistical methodology

Introduction

In the middle of the seventeenth century, a remarkable confluence of scientists, mathematicians, and philosophers laid the foundations for the theory of probability and formulated new philosophical underpinnings for the justification of claims to knowledge. These individuals knew one another, posed problems as challenges to one another, and criticized and defended the work of one another. Although investigations in probability had been conducted for well over two hundred years before, Fermat and Pascal were credited (by many historians of probability) with its mathematical development. Although many modern philosophical problems had been addressed by Aristotle, Socrates, and Protagoras, the interplay between probability and philosophy did not begin in earnest until the end of the seventeenth century and did not give birth to what Stigler (1986) called the infant discipline of statistics until 1900.

One reason for this fairly long dalliance is that it was not clear how the information provided by a probabilistic analysis could warrant knowledge claims, claims that at the time required justification as certain and true. Only slowly did probable knowledge get recognized as having any veracity, and this on a secondary level as opinion or belief. By the end of the eighteenth century, philosophers began to view even the possibility of acquiring certain knowledge of the real world as uncertain at best. It was only in the middle of the nineteenth century, when the philosophical focus shifted from the justification of the source of scientific knowledge to the validity of the methods of science, that the true romance between

Ronald C. Serlin is Professor in the Department of Educational Psychology at the University of Wisconsin-Madison. He teaches an introductory sequence in statistics, as well as courses in nonparametric statistics, multivariate statistics, and the philosophy of science and statistics. He won a University of Wisconsin teaching award, and he served two nonconcurrent terms as department chair. Email him at the following address: rcserlin@facstaff.wisc.edu

probability and philosophy blossomed in the testing of scientific theories.

This relationship continues to flourish, and the occasional disagreements are healthy, for "statistics requires a dynamic balance between its philosophical underpinnings and its practice to remain vital" (Kadane, 1976, p. 735). In order better to understand this balance and to maintain and strengthen the vitality of the applied and theoretical aspects of modern statistics, it will be helpful to examine the history of probability and its joint effort with philosophy of science. Such study will encourage researchers in statistical theory and methods to focus on problems whose solutions are essential to the continued health of the scientific enterprise, it will allow those researchers to avoid repeating mistakes of the past, and it is hoped that it will engender an appreciation for the incredible insights and magnificent oversights of our scientific forebears.

As Stigler wrote (1986), "the advances in scientific logic that took place in statistics before 1900 were to be every bit as influential as those associated with the names of Newton and Darwin" (p. 361). Indeed, even though Newton dabbled in probability theory, and Darwin's indirect affect on statistics through his cousin, Francis Galton, is well known, less well known perhaps are Newton's and Darwin's influence on philosophers of science and statistics. An understanding of these kinds of mutual influences of statisticians and philosophers may help to limn modern statistics in a new yet joyously familiar way, "...a recognition, the known appearing fully itself, and more itself than one knew" (Levertov, 1961).

Origins of Probability Theory

According to Walker (1927), the foundations of the theory of probability were laid by Blaise Pascal and Pierre de Fermat in 1654 in response to two questions asked of Pascal by Antoine Gombauld, the Chevalier de Mere, Sieur de Baussay. As with many, if not most, scientific advances, the work of Pascal and Fermat culminated the efforts of other scientists and mathematicians that had been accruing over a period of hundreds of years. Pascal and Fermat were first brought together through the auspices of Pierre de Carcavi and Marin Mersenne. Mathematicians, including Pierre Gassendi, Pierre de Carcavi, Gilles Roberval, Rene Descartes, and Blaise Pascal's father, Etienne, met at Mersenne's house once a week. Etienne introduced Blaise to the Mersenne Academy when Blaise was fourteen years old. Carcavi brought his friend Fermat, with whom he served in parliament in Toulouse, into correspondence with Mersenne and the others in 1636, and he suggested that Etienne and Roberval write to Fermat regarding their questions into methods of integration and centers of gravity. When Descartes criticized (erroneously) Fermat's method of finding tangents, it was Etienne and Roberval who defended him. Carcavi also first put Fermat and Blaise Pascal in touch with one another (David, 1962).

One of the questions that de Mere asked, known as the problem of points, concerned the fair distribution of stakes between two players when a game they were playing was interrupted mid-contest. The problem of points had been solved more than 250 years beforehand in some works by Antonio de Mazzinghi from around 1400 (Kiernan, 2001). The first time that the problem appeared in a mathematical work, it was solved incorrectly by Pacioli in 1494 (David, 1962; Kiernan, 2001). Cardano, who offered his own solution in 1539 (four years before Copernicus published his heliocentric theory!), referred to Pacioli's error as one that a child should recognize.

Unfortunately, Cardano's solution was wrong. In 1556, Tartaglia again took up the problem of points, commenting that Cardano's solution didn't make sense. Kiernan (2001, p. 181) notes that Tartaglia's answer was "way off", as well. Peverone in 1558 also attempted to solve the problem and failed, but according to David (1962), M. G. Kendall called this one of the near misses of history. It was not until Pascal and Fermat discussed the problem in a series of letters during the summer of 1654 that a correct solution was again found. This time the problem of points was solved in three different ways, one by Fermat using the enumeration of all cases, one by Pascal that used the process of recursion, and a second solution by Pascal using his arithmetic triangle. (The use of a triangular array such as Pascal's triangle to determine binomial

coefficients appeared in works by Chu Shih-chieh in 1303, Apianus in 1527, Stifel in 1545, and Tartaglia in 1556. According to David, 1962, Fermat dealt with it in 1636, which is perhaps the reason that Fisher has referred to it as Fermat's triangle.)

The second question posed by de Mere and solved by Fermat and Pascal dealt with probabilities associated with dice. He asked Pascal (and Roberval) why the probability of throwing at least one six in four rolls of a fair die was in the ratio 671 to 625, whereas the probability of obtaining at least one pair of sixes in twenty-four rolls of two dice was less than 0.5. Because the expected number of sixes rolled in four rolls of a single die is the same as the expected number of pairs of sixes in twenty-four rolls of two dice, the unequal probabilities that de Mere discovered led him, according to Pascal, to think he had found a "falsehood in the theory of numbers" and that "Arithmetic is self-contradictory" (cited in David, 1962, p. 88-89). That de Mere was able to distinguish empirically between two probabilities whose true values are 0.4914 and 0.5177, concluding that the former was less than 0.5, indicates that he was an assiduous gambler and note-taker.

Dice of reasonable quality are known to have existed since about 3000 B.C., used chiefly at the time in religious rites (David, 1962). A complete enumeration of the various outcomes on three dice appeared in a thirteenth century poem attributed to Fournival (David, 1962), and a 1477 commentary by Libri on Dante's Divine Comedy contains the first indication of the probabilities of various throws in a three-dice game of hazard (Todhunter, 1865). Cardano, however, possibly in concert with Ferrari, introduced in about 1526 (published posthumously in 1663) "the idea of combinations to enumerate all the elements of the fundamental probability set" and noticed that if all elements are equiprobable the ratio of favorable to total numbers of cases gives a result "in accordance with experience" (David, 1962, p. 58).

From this, David (1962) concluded that Cardano was the first mathematician to correctly calculate a theoretical probability. Unfortunately, Cardano was incorrect in his solution of what was essentially de Mere's second question. Galileo also took up the subject of dice games and published a fragment on them in around 1620 (David, 1962). His benefactor, to whom Galileo was Mathematician to his Serenest Highness, Cosimo II of Tuscany, had posed a problem that had been solved by Cardano and that was similar to that posed by de Mere: Why, in the throwing of three dice, is the number of partitions of 9 and 10 the same, though their probability in practice was not equal, with 9 being the less probable (David, 1962)? (His Serenest Highness was almost as discerning as de Mere, being able to distinguish between probabilities of 0.116 and 0.125.)

We can see that the topics addressed by Pascal and Fermat had a long history before the summer of 1654. Nevertheless, as Todhunter (1865) commented, "neglecting the trifling hints which may be found in preceding writers, we may say that the Theory of Probability really commenced with Pascal and Fermat" (p. 20). And yet, this work was never published by either Pascal or Fermat, though both desired that it be published.

It was Christian Huygens who incorporated their work into a small tract published in 1657, the first printed work on games of chance (Walker, 1929). Huygens learned the problem of points from one of Carcavi's friends (David, 1962). After Huygens solved the problem and sent his solution to Roberval, Carcavi sent Huygens the outlines of the discussion of the problem between Fermat and Pascal, and he later sent Fermat's solution to Huygens, which turned out to be the same as Huygens'. Fermat posed even more difficult problems to Huygens, which he solved and incorporated into his tract (David, 1962). According to David (1962), if one says that "the real begetter of the calculus of probabilities is he who first put it on a sound footing" (p. 110), then one should look to Huygens, Lord of Zelem and of Zuylichem, "the scientist who first put forward in a systematic way the new propositions..., who gave the rules and who first made definitive the idea of mathematical expectation". For nearly fifty years, Huygen's work (in Latin) was the unique introduction to the theory of probability (David, 1962). Todhunter (1865) attributes a 1692 English translation of Huygens' tract to John Arbuthnot.

Newton was familiar with Huygens' writings (David, 1962). With the arrival of The Great (bubonic) Plague (1664-65), Trinity University was closed, and Newton retired to Woolsthorpe for two years to invent calculus, discover the universal law of gravitation, and prove experimentally that white light is composed of all colors. Newton's *Principia* was presented to the Royal Society in 1686 and published in 1687 (printed at Edmund Halley's expense), thirty years after Huygens published the work of Pascal and Fermat. And in 1693, Newton solved what was essentially de Mere's dice problem in response to a query by Samuel Pepys, thus revealing what David (1962) described as at least elementary knowledge of probability theory.

Certain Knowledge

Probability theory has clearly long been of interest to gamblers. As Bellhouse (1993) noted, "familiarity with probability theory can enhance the strategy of play." Putting the parentage of the theory aside, one must wonder, given that Pascal and Fermat's theory culminated well over one hundred years of work on probability, why the methods of probability were not beginning to be incorporated into the scientific pursuit of knowledge. David (1962) opined, "At a time when it was still possible for an able mathematician to take all knowledge for his province, moreover when dicing, and gambling with annuities, were practiced as assiduously in England as anywhere else, it is indeed strange that not only Newton but nearly the whole of the English school showed no interest in them" (p. 124-125).

David (1962) suggested that the introduction of probability into science did not come before the Renaissance "because the philosophic development which opened so many doors for the human intellect engendered a habit of mind which made impossible the construction of theoretical hypotheses from empirical data" (p. 26). One or another form of Aristotelianism was dominant at the beginning of the seventeenth century (Garber, 1995). And yet, even late into the Renaissance, during a period in which Newton seemed to have obtained hypotheses from data (despite his *hypotheses non fingo* claim to the contrary), probability had yet to enter the scientific arena.

One possible reason for this late entry of probability into scientific method is that in the middle of the seventeenth century, and through the middle of the nineteenth century, knowledge was defined as certified true belief. Indeed, even Pascal claimed that he was not satisfied with the probable, seeking instead the sure (Watkins, 1978). At the heart of this epistemological view, according to Suppe (1977), was the argument that S knows that P if and only if (a) P is true, (b) S believes that P, and (c) S has adequate evidence for believing that P. From the late sixteenth through the early twentieth centuries, natural philosophers were preoccupied by systematic methods for discovering knowledge (Mulaik, 1987). In this regard, the justification clause (c) was satisfied only by finding a demonstrably incorrigible base knowledge consisting either of the intuitionist Descartes' a priori clear and distinct ideas or by the sense data of inductivists such as Bacon and Gassendi.

Greek philosophers recognized that the senses can deceive us. For example, atomists such as Democritus believed the world to be made from tiny entities known as atoms whose action on the senses cause us to experience smell and heat, for example. Yet, as the atoms have no smell or heat, the world of appearance is illusory (Mulaik, 1987). For Descartes, whom Peirce called "the father of modern philosophy" (Peirce, 1868), the broadest aspects of nature are understood by deduction from incorrigible first principles, which are grounded in pure reason (Salmon, 1966).

So committed to certainty was Descartes that in his *Discourse on Method* of 1637 he claimed as false all that was only probable. According to Cartesianism, the world is full of an infinitely divisible matter, reason dominates, and philosophy is based on his own clear and distinct perceptions (Garber, 1995). For example, as Descartes wrote in his *Meditations* (1642), "Now it is manifest by the natural light that there must at least be as much reality in the efficient and total cause as in its effect. For, pray, whence can the effect derive its reality, if not from its cause?" Salmon wonders how the intuitionist Descartes, a man who could not be certain that 2+2=4 or that he had hands unless he

could prove that God is not a deceiver, found it impossible to conceive of the falsity of the foregoing principle.

Descartes prepared his *Meditations* in Holland in 1640. Huygens transported it in manuscript form to Mersenne, who solicited responses from "learned men who would take the trouble to scrutinize them" (Descartes, cited in Joy, 1995, p. 431). Among those who contributed were Hobbes, Gassendi, and Mersenne, himself. According to Agassi (1975), Gassendi asked why one would deduce "I think, therefore I am?" Why not "I walk, therefore I am?" Descartes understood the point and agreed that if one walked, one necessarily existed. But he could not be sure that he walked; he could be sure that he thought, and that is why he preferred his "Cogito". He didn't doubt the validity of Gassendi's inference, he only doubted the truth of the premise that he walked. (Agassi misattributed this Fifth Objection to Hobbes, who actually wrote the Third.)

Gassendi was an empiricist. For him, experience dominates, and philosophy begins with our sensations of a public world; this world is made up of atoms and a void, and he attempted to reconcile Epicurean atomism in a way that was more congenial to the Church. In rejecting Aristotelianism, he, like Descartes, adopted the mechanist philosophy's premise that physical phenomena could be described fully in terms of matter and motion. He also believed that our senses can fool us, which caused him to formulate a kind of moderate skepticism that influenced Locke, Peirce, and others.

For other empiricists, like Bacon, the justification of scientific theory is based on its ability to explain experimental results. Until Bacon, logic as described in Aristotle's *Organon* (Greek for "tool") was deductive. What was needed was a method that abandoned Aristotelianism's approach that began with hypotheses and deduced truths from them (Mulaik, 1987). Bacon introduced his inductive logic in his *Novum Organum* (Latin for "New Tool") in 1620. According to Bacon's doctrine (Lakatos, 1978), a discovery is scientific only if it is guided by facts through a method of induction "that would begin without hypotheses or speculations, systematically interrogate nature, and move to ever more general truths by means of an automatic procedure or algorithms" (Mulaik, 1987, p. 273). The scientist starts by clearing his mind of theory (bias), and nature will then make itself known. For Bacon, science is an experimental enterprise through which one investigates phenomena in controlled circumstances. Bacon's method of eliminative induction includes the logical insight that affirming instances do not provide evidence for inductive generalizations, whereas negative instances do provide disconfirming evidence (Mulaik, 1987). Bacon, apocryphally, died of pneumonia that developed while he was investigating refrigeration by stuffing a chicken with snow.

Although Bacon's *Novum Organum* of 1620 preceded Descartes' *Discourse on Method* by seventeen years, Descartes' philosophy was dominant at the time of Newton's *Principia*. According to the justificationist standards of the day, then, Newton's theory was non-knowledge (Lakatos, 1978). Newton's theory was not proved in the Cartesian sense, because it was not derived from Cartesian metaphysics. Newton instead proposed that propositions required only an empirical-experimental and not a rational-metaphysical proof (Lakatos, 1978). Because of the extraordinary success of Newton's theory, "for 200 years after Newton no one could advocate the use of hypotheses without an uneasy backward glance" (Medawar, 1974). This, despite inductivism having suffered what should have been severe setbacks at the hands of Locke, Hume and Kant.

Probable Knowledge

The beauty and power of Newton's mathematical approach to physics clearly had an effect on John Arbuthnot, who wrote in 1692, "There are very few things which we know; which are not capable of being reduc'd to a Mathematical Reasoning; and when they cannot, it's a sign our Knowledge of them is very small and confus'd" (Stigler, 1986, p. 1). Arbuthnot implemented a binomial test in 1710 to examine "the constant regularity observ'd in the births of both sexes," (Stigler, 1986, p. 225), and he is often credited with publishing the first statistical test. Fisher, however, attributed the first published significance test to de Moivre in 1718, and Barnard stipulated that the first published

test was due to Daniel Bernoulli in 1734 (Bennett, 1990, p. 23-26). Regardless of which test is deemed to have been the first, it is clear that the eighteenth century held promise for great discoveries in probability and statistics. Some of the early discoveries in probability and statistics were important to philosophers, as well. Jacob Bernoulli developed the theory of permutations and combinations and contributed the weak law of large numbers, the theorem that with an increasing number of observations, the probability increases that an estimator will lie within any specified distance of the true value.

According to Stigler (1986), at least five Bernoullis worked on probability, writing "So large is the set of Bernoullis that chance alone may have made it inevitable that a Bernoulli should be designated father of the quantification of uncertainty" (Stigler, 1986, p. 63). Jacob Bernoulli and philosopher Gottfried Leibniz are known to have composed twenty-one letters to one another, although one may not have been sent (Sylla, 1998). Leibniz may have first learned of Jacob's work in probability from Jacob's brother, Johann, with whom Jacob was not speaking. In a letter written in 1697, Leibniz spoke of the "need for establishing on firm foundations an art of measuring degrees of proofs" (Sylla, 1998, p.48). And after the publication in 1713 of Jacob Bernoulli's *Ars Conjectandi*, accomplished eight years posthumously by his nephew Nicholas because of the rift between brothers, Leibniz noted that the probabilities of obtaining an 11 and a 12 in rolling two dice are equal.

John Locke is considered to be the father of British empiricism, and he is perhaps the first major philosopher to discuss probable knowledge as a somewhat tenable, "second-rate way of becoming cognitively aware of the nature of the world" (Owen, 1993, p. 38). For Locke, probable knowledge is faith or opinion. Owen noted that Locke and other non-Cartesians stood at a junction between the old and new ways of looking at the world. Locke's account "recognizes the limitations of knowledge, rather traditionally conceived, but looks ahead in allowing its rational supplementation by probable conjectures" (Owen, 1993, p. 39).

In his 1690 *An Essay Concerning Human Understanding*, Locke sought to support Bacon's empiricism by arguing that knowledge can not have a component based on innate ideas. He argued that if knowledge is not received through the senses, then the mind at birth must have some kind of intellectual ability, at least in applying the concepts of logic (Clark, 1957). Instead, he felt that a person enters the world with a mind that is a blank slate. There are only two sources of ideas, sensation and reflection. For Locke, complex ideas are formed out of the simple ones entering the mind through the mental activities of compounding, abstracting, and relating. By a method of analysis, Locke was able to trace back from complex ideas to the simple ones out of which they arose, but he could not find the simple idea from which the concept of substance came (Mulaik, 1987). Because of this, and because he argued that the certain qualities of objects, such as color and odor, exist only in the mind and are not representative of reality, we can not be certain that any of our ideas are representative of reality.

The case for the demise of inductivism was made well and irremediably in David Hume's *Enquiry concerning the human understanding* of 1748. Hume's objections to induction can be variously phrased. According to Harris (1992), Hume concluded that it is impossible to justify epistemologically that unobserved cases will resemble observed cases in some crucial respect. Because of this, neither certain nor probable knowledge can be justified.

Reichenbach (1951) discussed two theses put forward by Hume. In the first thesis, Hume makes clear the nonanalytic nature of induction by pointing out that we can very well imagine the contrary of the inductive conclusion. The possibility of a false conclusion in combination with a true premise proves that the inductive inference does not carry a logical necessity with it. Hume's second thesis is that induction cannot be justified by reference to experience--the inference with which we want to justify induction is itself an inductive inference (we believe in induction because induction has so far been successful), and so we are caught in circularity. Russell (1945, p. 672) stated Hume's conclusion as, "We cannot help believing, but no belief can be grounded in reason." Of Hume's conclusion, Russell (1945) exclaimed, "It is therefore important to discover whether there is

any answer to Hume within the framework of a philosophy that is wholly or mainly empirical. If not, there is no intellectual difference between sanity and insanity. The lunatic who believes that he is a poached egg is to be condemned solely on the ground that he is in a minority" (p. 673). It would seem that as of 1748, unless arguments could be mounted against Hume's attack, inductivism was dead. Yet, it lived on, because of the success of Newton's theory.

Expanding on the work of Jacob and Nicholas Bernoulli, De Moivre published the first appearance of the normal curve in 1733 (Stigler, 1986). And in 1763, Bayes' Theorem was published posthumously by Richard Price, who presented it to the Royal Society. Fisher (1956) thought Bayes was reluctant to publish his work because Bayes felt that his postulating a uniform prior distribution might be considered disputable. Price, according to Gillies (1993), was strongly influenced by Hume's criticisms of induction and thought that Bayes' Theorem could be used to resolve the problems raised by Hume by making generalizations probable, rather than certain (this despite Hume's injunction against such a possibility).

Synthetic *a priori* Knowledge

The first major intuitionist response to Hume's empiricist attack was due to Kant, who wrote *Critique of pure reason* in 1781, according to Reichenbach (1951), "with the intention of saving scientific knowledge from the annihilating consequences of Hume's criticism." Kant, who in his preface to the *Critique* compared his work to that of Copernicus, made clear two distinctions among types of propositions. First, he distinguished between analytic propositions (true virtually by definition, such as the statement "All bachelors are unmarried") and synthetic propositions (those that inform us about a fact, such as observations, and add to our knowledge). Second, he distinguished between *a priori* propositions, those which have a basis other than experience, and *a posteriori* (or empirical) propositions, needing observational evidence to determine their truth. He posited that objects conform to the conditions set forth by the mind, that whereas the senses provide the subject matter, the mind imposes the form of thought. Rather than the mind being a Baconian blank slate, Kant specified what he called the categories of thought as the *a priori* equipment for thinking. He felt that by showing that the axioms of Euclidean geometry were synthetic and yet known *a priori*, he could establish the incorrigible basis that justified Suppe's clause (c) mentioned earlier. It would seem, then, that at this point, intuitionism held the upper hand, due to Hume's crushing blow against inductivism and to Kant's intuitionist argument that Euclidean geometry was synthetic and yet known a priori.

The nineteenth century saw major upheavals in science and philosophy. As described by Reichenbach (1951), "Ever since the death of Kant in 1804 science has gone through a development, gradual at first and rapidly increasing in tempo, in which it abandoned all absolute truths and preconceived ideas." Lagrange introduced the method of least squares in 1805, and in 1809 Gauss addressed the same problem but couched it in probabilistic terms (he also claimed priority for the method of least squares, claiming he had used it since 1795 - Stigler, 1986).

Laplace contributed the central limit theorem in 1810, inverse probability and the principle of insufficient reason in 1812. His definition of probability was as a state of mind (Fisher, 1956; Epstein, 1977), whereas Bayes seems to have used a frequentist definition (Fisher, 1956). The definition of probability as the limit of a frequency was due to Poisson in 1837. According to Epstein (1977), the theory of probability is more indebted to Laplace than to any other mathematician; indeed, Stigler (1986, p. 122) claims that Laplace's work brought about "a truly Copernican revolution in statistical concept." The Gauss-Laplace synthesis brought together two lines - the combination of observations and the use of probability to make inferences - into a coherent whole that was widely disseminated through the middle of the century (Stigler, 1986).

But Gauss, along with Bolyai and Nikolai Ivanovich Lobachevsky, called the Copernicus of geometry by English mathematician William Clifford (Bell, 1937), made a discovery that had far greater philosophical import - the discovery of

non-Euclidean geometry. Lobachevsky's publication appeared in 1829-30 and Bolyai's in 1832. Gauss claimed to have obtained similar results earlier but did not publish because, according to Gillies (1993, p. 80), "he was 'afraid of the clamour of the Boeotians.' Boeotia was a region of ancient Greece whose inhabitants were considered by the Athenians to be stupid and uncultured" (p. 80). The arrival of non-Euclidean geometry showed that Kant's implication that humans could never conceive of non-Euclidean geometries was untenable. Despite this, Kant's impact was strong and lasting.

Descriptive Knowledge

Burtt (1924) saw elements of positivism in Galileo's work, and Burtt cited Brewster's claim that Newton was the first great positivist. The founder of positivism in its 19th century form was Auguste Comte. Comte's *Cours de philosophie Positive* was completed in 1842. Comte is also known as the founder of sociology. Positivism was Comte's response to the upheavals in society and to Laplace's "scientifically reasoned deterministic interpretation of the universe" (Epstein, 1977, p.7). It was Comte's hope that science could be turned into a religion, "in which the great philosophers and scientists took the place of the Christian saints, and an organized devotion to the cause of humanity was substituted for the worship of God" (Fuller, 1938, p. 384). According to Comte, there are three stages in the history of thought: 1) a theological stage, explaining the universe in terms of the purposes of deities; 2) a metaphysical stage, explaining in terms of abstract principles which are personified; and 3) a scientific stage, in which uniformities in nature are described without reading any evidence of purpose or design or consciousness into them. The meaning of terms are referred to what is found in experience.

Positivists eschew metaphysics and refrain from explanation in physics. Science organizes knowledge using laws that are merely descriptions, approximate at that, of the patterns in which phenomena occur, and science gives us the power of prediction. Bradley (1971) paraphrased Martineau in saying it is strange that something so negative should be called positivism.

Fortunately, although an actual Religion of Positivism was started, with priests, rituals, and baptisms, most of Comte's excesses in this direction were ignored. Comte's positivist heir was physicist Ernst Mach, who was ecumenical in his influences, including Hume, Kant, and Darwin (Cohen, 1970, p. 127). According to Cohen (1970), Mach "apparently succeeded in combining a Kantian appreciation of the active, even constitutive, role of the mind in generating science with a scientific, which is to say, empirical-biological, theory of the origins and functions of the mental life" (p. 156). For Mach, "not knowledge attained, but the method of attaining it, could be certified" (Cohen, 1970, p.129).

Mach, like Comte, was an instrumentalist and felt that laws were mere descriptions of nature. Mach, however, did not completely do away with theories (as opposed to laws), as long as they were testable. Mach's positivism differs from Comte's in that nothing was "more foreign to Mach than the tendency towards absolutism which finally disfigured both the philosophical and the human image of Comte" (von Mises, 1970, p. 266). Even by the turn of the twentieth century, physicists such as Plank and Einstein, although influenced greatly by Mach early on, began to turn against positivism.

Conjectural Knowledge

William Whewell, who coined the word 'scientist' (as well as 'anode' and 'cathode' for Faraday and the words 'physicist', 'eocene', 'miocene', and 'pliocene' - Medawar, 1974) upon the request of the poet Samuel Taylor Coleridge in 1833, tried to reformulate the problems of the philosophy of science in a Kantian way (Wettersten, 1993), while not relying on Kant's fixed a priori categories. He attempted to "explain the facts of the growth and stability of science without appeal to induction, which he saw to be useless" (Wettersten, 1993, p. 482). In his *Novum Organum Renovatum* of 1858, Whewell considered induction to be "the representation of facts with principles" (Wettersten, 1993, p. 497), a notion that will be seen in the pragmatacist philosophy of Charles

Sanders Peirce, and not the Baconian induction from facts to generalizations. He showed that neither empiricism nor intuitionism, including Kant's, could account for the growth of scientific knowledge; instead, both experience and intuition were needed. He gave importance to independent tests and to new predictions, and he claimed that science needs guesses (Medawar, 1974 noted that Whewell also used the phrase 'felicitous strokes of inventive talent' when a more formal phrase than 'happy guesses' was required.) As Medawar (1974, p. 281) explained, "To say that Einstein formulated a theory of relativity by guesswork is on all fours with saying that Wordsworth wrote rhymes and Mozart tuneful music. It is cheeky where something grave is called for to explain how scientists discover true principles." According to Wettersten (1993, p. 506), Whewell's theory makes clear that "even if we start with poor guesses and treat them critically we can come to the truth: there are many paths to the truth but only one goal'. We see then that Whewell's approach is essentially deductivist and that the process consists above all in criticism. In this, Whewell is a direct predecessor to Karl Popper's philosophy of conjectures and refutations (Wettersten, 1992).

According to Reichenbach (1951), "the turning point in the history of logic was the middle of the nineteenth century, when mathematicians like Boole and de Morgan undertook to set forth the principles of logic in a symbolic language." Peirce, a mathematician and logician by training, carried on this work. It was not until Boole, DeMorgan, and Peirce mathematically overhauled traditional formal logic that the logic of probability was put on a more scientifically useful basis (Wiener, 1972). That Peirce was a frequentist could have been due to Boole's strong criticism in 1854 of the postulate of which Bayes was so chary. Like Whewell, Peirce was heavily influenced by Kant. He claimed that he read Kant's *Critique of Pure Reason* two hours per day for three years, and he named his philosophy 'pragmatism' in honor of Kant, whom he called The Philosopher. He did not use the term practicalism, because in Kant pragmatism and practicalism are virtually polar opposites (Buchler, 1939).

Pragmatic means empirical or experimental, whereas Kant's notion of practical laws are given purely a priori. Indeed, so often were these terms misunderstood that Peirce threatened to call his philosophy pragmaticism, a term he felt was so ugly that it wouldn't be kidnapped.

According to Wiener (1972), the great difference between the American pragmatists and Kant is their denial that over and above contingent pragmatic belief are the purely rational, necessary, and absolute ideas of Kant's transcendental philosophy. The purpose of inquiry, wrote Peirce, is to enable us to pass from a state of doubt to a state of belief. Despite his high regard for Kant, Peirce's philosophy differed from that of Kant. For example, whereas Kant considered mathematics to be synthetic and yet true a priori, Peirce held that mathematics and logic are not synthetic (Buchler, 1939).

He also provided his own version of Kant's categories, writing of them that in making their character unchangeable, Kant was hostile to the spirit of empiricism. Because of the constant nature of Kant's categories, Kant's epistemology formed a closed system. But Peirce, having the benefit of Darwin's *Origin of Species* of 1859, provides an adaptive mechanism behind his categories. Peirce attempted to convert the Darwinian ideas of chance variation and natural selection into the idea of an evolution of the mind by means of a logical competition among thoughts, which eliminates ideas not fit to stand for the truth fated to be discovered by those who investigate. It was the nonevolutionary character of the old forms of a static empiricism and a rigid *a priori* intuitionism that engaged the pragmatists.

Peirce was a fallibilist, extending the views of Gassendi and Locke in a most thorough way. "I will not," he wrote, "admit that we know anything with *absolute certainty*. It is possible that twice two is not four" (Peirce, 1958, p. 64). Although he felt that the notion of certain knowledge is absurd for a variety of reasons, there were two main reasons underpinning his fallibilism. First, all claims to knowledge are criticizable and only held conditionally, for there is no ultimate inductivist or empiricist basis that can stop the respective infinite regress in the

justification of the claims. And second, he felt that no theory was true, able to satisfy all features of the facts. In terms of Newton's law of gravity, he pointed out that if, instead of inverse square attraction, the exponent of the distance between bodies was 2.000001, there would only be a minor consequence observable in the orbits of the planets, resulting in only slight discrepancies in estimated planet masses (Peirce, 1958).

Peirce (1878) classified all inference as either deductive (or analytic) or synthetic, which he subdivided into induction and hypothesis. (One difficulty encountered in reading Peirce results from his using 'hypothesis', 'retroduction', and 'abduction' for the same synthetic inference. In addition, Peirce delineated several types of induction.) Deduction is a syllogism in which the truth of a rule and a case is transmitted to the result, and conversely from the falsity of the conclusion, the falsity of the premise follows. In induction, we infer from a number of cases that the same thing is true of a whole class. Peirce showed that an induction is the inverse of a deductive syllogism, so that from the case and the result, the rule is inferred. As an example (Peirce, 1878), from the deduction:

Rule: All the beans in the bag were white.

Case: These beans were in the bag.

Result: These beans are white.

we can obtain the induction:

Case: These beans were in the bag.

Result: These beans are white.

Rule: All the beans in the bag were white.

Hypothesis infers the case from the rule and the result:

Rule: All the beans from this bag are white.

Result: These beans are white.

Case: These beans are from this bag.

Peirce described the scientific method in terms of these three modes of inference in the following way (Peirce, 1958):

> Accepting the conclusion that an explanation is needed when facts contrary to what we should expect emerge, it follows that the explanation must be such a proposition as would lead to the prediction of the observed facts
>
> A hypothesis then, has to be adopted, which is likely in itself, and renders the facts likely. This step of adopting a hypothesis as being suggested by the facts, is what I call *abduction*.
>
> [T]he first thing that will be done, as soon as a hypothesis has been adopted, will be to trace out its necessary and probable experiential consequences. This step is *deduction*. (p. 122).

An abduction for Peirce is an explanation.

The third step in the process involves induction (Peirce, 1958):

> Having...drawn from a hypothesis predictions...we proceed to test the hypothesis by making the experiments and comparing those predictions with the actual results of the experiment.
>
> This sort of inference it is, from experiments testing predictions based on a hypothesis, that is alone properly entitled to be called *induction*.
>
> Induction...is not justified by any relation between the facts stated in the premisses and the fact stated in the conclusion...But the justification of its conclusion is that that conclusion is reached by a method which, steadily persisted in, must lead to true knowledge in the long run. (p. 124-125)

Peirce distinguished two major types of valid induction (there is actually a third type that Peirce called the Pooh-pooh argument, but enough said). The first, quantitative induction, involves the ascertainment of a ratio in the population from samples. Through this type if induction, we can attain moral certainty of the population value, by which Peirce means a probability of 1 based on Bernoulli's results concerning the probability that the sample value lies within certain limits of the population value. "Of course," he wrote, "there is a difference between probability 1 and absolute certainty" (Peirce, 1958, p. 131). The second type of induction Peirce called qualitative induction, from which the most that can be said is that there is no reason yet for giving up the hypothesis. Of this second type, Peirce (1958) wrote, "the only justification for this would be that it is the result of a method that persisted in must eventually correct any error that it leads us into" (p. 134).

Peirce claimed for induction a trustworthiness because of the manner of proceeding (Buchler, 1939). The concept of a probable argument referred to a class of arguments, and an induction belongs to the class of all inductions. Saying an induction was probable meant that the majority of inductions were successful. "[T]hat real and sensible difference between one degree of probability and another...is that in the frequent employment of two different modes of inference, one will carry truth with it oftener than the other" (Peirce, 1878).

Neither qualitative nor quantitative induction and the associated probabilities of success involves the probability that a generalization itself is true. According to Buchler (1939), "After 1883 Peirce does not even regard induction as 'probable'...but rather as not probable at all" (p. 251). Peirce said that talking about the probability of a law was nonsense, as if universes were as plentiful as blackberries, and we could pick one. This later view reflects Peirce's distinction between two types of probability, the empirical probability associated with ratios or with the class of inductions and what Peirce called conceptualistic probability that is not strictly a probability, but is instead only a sense of probability (Buchler, 1939).

As with Whewell, Peirce emphasized that potential explanatory hypotheses are formulated as guesses. For Peirce, as with Mach, the force of scientific reason lies in its methods. "[T]he *method of methods*, is the true and worthy idea of the science" (Peirce, 1958, p. 44). Science is rational, according to Peirce (1958, p. 49), where "...'rational' means essentially self-criticizing, self-controlling and self-controlled, and therefore open to incessant question." And rather than leading to the probability that the inductive inference itself is true, the ability to draw valid conclusions lies with the probability of correctness of its inductive method, "the relative frequency with which this class of inferences is found to yield true conclusions" (Buchler, 1939, p. 233).

Unprovable and Improbable Knowledge

By the end of the nineteenth century, the philosophical focus was on American Pragmatism and Machian positivism. Both Galton and Pearson were Machian instrumentalists, which would at least partly explain Pearson's emphasis on fitting data to his own system of curves. The continuation of Mach's doctrines fell to the logical empiricists. The response of Russell and the Vienna Circle philosophers was to search for an empirical basis and an inductive logic. Realizing that justifying an inductive principle on the basis of observation would lead to an infinite regress - to justify it would require inductive inferences - Russell advocated accepting the principle of induction on the ground of its intrinsic evidence (Gillies, 1993), that is, on an a priori basis. But even if we accepted *a priorism* as a justification of an inductive principle, the positivists' search for an empirical basis was doomed to failure, as shown by Duhem, who advanced two theses against inductivism. One of these, afterwards to become known as the Duhem-Quine thesis, will be discussed later.

The other thesis shows that all observations are theory-laden. According to Agassi (1983), the claim that empirical evidence has a theoretical bias was recognized by Bacon and Galileo; if one has a theory, it biases perception. This led to Bacon's request that

scientists first make observations with no theory in mind. Galileo realized, of course, that this would result in "just a heap of observations" (Agassi, 1983, p. 10), and he was convinced that geometry, based on *a priori* intuitions, must precede facts. This led to Kant's argument against empiricism, and Whewell, influenced by Kant, deduced that all data are interpreted, either on the basis of theory or of *a priori* intuitions. Therefore, trying to prove a theory inductively ultimately requires proving a theory from a theory, which is impossible. All one could conclude on this basis is that the theories involved are consistent. Thus, the theory-ladenness of observations meant that theories could no longer be hoped to be proved from an incorrigible basis.

It was still felt, however, that although theories may not be provable, they still could be disproved, or falsified, a view that flies in the face of the Duhem's second thesis, which states that an experiment can never condemn an isolated hypothesis but only a whole theoretical group. Underpinning this thesis is the realization that no theory can specify any observable consequences. Rather, it requires the conjunction of the theory, initial conditions, and auxiliary hypotheses. Thus, there can not be such a thing as a crucial experiment, on the basis of which a theory is falsified and dropped, because an observation contrary to prediction can only condemn the collective and not any individual part. Quine (1951) concluded that any statement can be held to be true, if we make enough adjustments elsewhere in the system. Thus, not only did the theory-ladenness of observations make theories unprovable, the Duhem-Quine thesis makes them undisprovable. So positivists had to fall back on the hope that theories could at least be shown to be probable.

Neyman and Pearson (1933) and Fisher (1935) approached these issues from different perspectives, and certainly different from the probabilist approach of Jeffreys (1939). For probabilists, theories have different degrees of probability (Lakatos, 1978). Scientific honesty then consists in uttering only highly probable theories, or the probability in light of the evidence. But Ritchie (1926) showed that the probability of any inductive generalization is zero, and Lakatos (1978) points out that in the early 1940's, Carnap found that the degree of confirmation of all genuinely universal propositions was zero. So not only can no theory be proved or disproved with certainty, but theories are also equally improbable. This, then, was finally the end of positivism.

Criticism and Knowledge

Popper, in his *Logic der Forshung* in 1934 (Popper, 1959), attempted to address the issues that have been raised, especially Hume's skepticism, the theory-ladenness of observations, and the inability to condemn a hypothesis in isolation. In his solution, we can see much of what was good in Hume, Kant, Mach, and especially Whewell and Peirce. Popper's view of knowledge is fallibilist, as was Peirce's, and for him method is fallible as well, as distinguished from Mach's view that method was certain. Indeed, Peirce's overall view of the inductive process is virtually indistinguishable from the conjecture-and-refutation model advocated by Popper (Wiener, 1972). Popper (1962) claimed that his method of conjectures and refutations had its origins in the writings of Kant. Popper never questioned Hume's indictment of induction; instead, he insisted there was no problem. Instead of an inductive principle, Popper advanced "the theory of the deductive method of testing, or as the view that a hypothesis can only be empirically tested--and only after it has been advanced" (Popper, 1959, p. 30).

Musgrave (1993) described Popper's solution to the problem of induction in the following way. Popper, he said, rejected the assumption that an ampliative hypothesis is reasonable if, and only if, it is justified by the evidence, if, and only if, the evidence shows it to be true or probably true. In this, it is not clear whether justifying beliefs refers to justifying the things we believe or providing a warrant for our believing those things. According to the classical argument, we are justified in believing something if, and only if, we can show it to be true or at least show it to be more likely true than not. Popper rejected this assumption, allowing him to endorse Hume's inductive skepticism while rejecting his irrationalism. To get from the skeptical thesis to the irrationalist thesis you also must assume that a belief is

reasonable if and only if it is justified. Popper rejected this also.

In Musgrave's (1993) view, Popper affirmed that some evidence-transcending beliefs are reasonable. The central claim of Popper's approach, said Musgrave (1993), is that an evidence-transcending belief is reasonable if, and only if, it has withstood criticism, including, where appropriate, attempts to refute it by appeal to evidence. When a prediction is falsified we will say that what we predicted was wrong, not that it was unreasonable to have predicted it. For any reasonable theory of reasonable belief, according to Musgrave (1993), must make room for reasonable beliefs in untruths. In short, Hume's criticism of induction applied to the search for a warrant for our beliefs, whereas in Musgrave's view, it does not apply to obtaining a warrant for our act of believing.

By contrast, according to the pancritical rationalism of Bartley (1984) and the comprehensively critical rationalism of Miller (2002), reflecting and extending the philosophy of Peirce and Popper, "neither beliefs nor acts of belief, nor decisions, nor even preferences, are reasonable or rational except in the sense that they are reached by procedures or methods that are reasonable or rational...Still less are beliefs, or decisions, or preferences ever justified" (Miller, 2002, p. 81). According to Miller (1982), the major difference between Popper's falsificationism and the justificationist philosophy of others is methodological, not epistemological.

Virtually all modern philosophers of science agree that certain knowledge can not be attained. Popper was the first to say outright that the attempt to attain certainty should not even be made. Miller (1982) pointed out that for justificationists, a hypothesis has to be confirmed, perhaps inductively, before it is admitted to science, and if it fails the tests, or is disconfirmed, or not confirmed at all, it is excluded from science. For Descartes, ideas that can not be justified by being reduced to clear and distinct ideas should be rejected, and anything that is accepted must be justified in this way. For Hume, any idea that can be justified by being derived from experience, the empiricist's only source of knowledge, should be accepted, and any idea that can not should be rejected (Bartley, 1984).

For Popper, as with Peirce, a hypothesis is tested only after it is admitted by being conjectured. There is a policy of "open admission", restricted only by the requirement that no hypothesis be admitted without there being some way to test it (Miller, 1982, p. 22). If the hypothesis passes a test, nothing happens, whereas if it fails a test, it is expelled. Because of the open admission policy, "it is of the greatest importance that the expulsion procedures should be brought into play at every possible opportunity...If we are seriously searching for the truth, we should submit any hypothesis proposed to the most searching barrage of criticism, in the hope that if it is false it will reveal itself as false" (Miller, 1982, p.23).

Criticism

One objection that could be raised regarding the critical rationalist methodology concerns the use of logic in a rational approach to science. Surely, this line of thinking would go, the principles of logic must be assumed to be true on an *a priori* basis. Are we not committed to an un-revisable logic, because logic itself can not be used to criticize logic? It is true that "critical argument...cannot be carried on without some system of logic. You cannot in this sense abandon logic and remain a rationalist" (Miller, 1994, p. 91). But the system of logic one uses can be criticized if the logical rules consistently lead to errors. Miller (1994) gives the example of a program written in FORTRAN that can be used to test the correctness of an operating system, even though the operating system is presupposed. Miller (1994) noted that it is "logic itself" (p. 91) that is supposedly assumed to be beyond criticism by critical rationalism. Yet, logic is involved in the critical argument in a particular formulation, at a minimum usually involving the principle of noncontradiction and the law of excluded middle, which might be right or wrong, and not in an unformulated way as logic itself. And whatever the particular formulation, it can certainly be criticized.

Does not the approach presuppose an inductive principle, such as the uniformity of nature or that the future will resemble the past, at least as far as specifying that we expect that

the laws we've discovered should work in the future? As Miller (1982) pointed out, "In order to provide genuinely interesting knowledge of the world inductivism needs to assume that there is some order and regularity in the world, whilst falsificationism requires only that there is some order and regularity in the world–but it does not need to make any sort of assumption to this effect" (p. 33). Miller went on to note that if there were no regularity, falsificationism would yield little, except the conjecture that there is no regularity. Hypotheses propose order, but if there is none, none will be found. They do not presuppose it.

As regards the reliability of a theory, no theory is reliable, in that Hume showed that without an inductive principle such as that the future will resemble the past, there is no logical way to infer that the theory will work in the future (or that it will fail). But if a theory is conjectured and stands up to severe testing, then it has not been discorroborated (a term used to emphasize the tentative nature of falsifications), and it may be tentatively classified as true; and one can *deduce* from the conjecture that various predictions will hold without relying on the uniformity of nature. As Miller (1980) wrote, "Whatever one calls them, Hume's problem simply does not arise for guesses" (p. 123). But, the issue might be pursued, if theories are unreliable, then why should any decisions be based on them?

Again, it seems rational to base a decision on a theory that has stood up to severe testing instead of one that has failed a severe test. As Miller (2002) pointed out, if one wants to avoid bad outcomes tomorrow, he can cross his fingers or he can try to be rational today. This does not mean, of course, that we can not hope that our favorite theories will continue to stand up to severe criticism. Radnitzky (1982) explained, "we have a subjective belief that the regularities described by a highly corroborated theory will also hold in the future. But this subjective belief is not granted any methodological significance" (p. 74).

Finally, the question arises as to how one could base a rejection of theory on the basis of experience if all basic statements are tentative. In this regard, Popper (1985) pointed to the well-known asymmetry between corroboration and rejection, namely that no matter how many confirmatory observations are observed, a theory can never be proved, whereas a single disconfirmatory observation can falsify (tentatively) a theory. Thus, as regards the observational basis, "No matter whether they are true or whether they are false, a universal law may not be derived from them. However if we assume that they are true the universal law may be falsified by them" (Popper, 1985, p. 185). Here the basic statements are conjectured to be true and are severely tested. "No falsification is conclusive," Miller (1982) wrote, "if only because all test statements are themselves fallible and open to dispute. But it would be incorrect to conclude from this that no hypothesis can be properly falsified... [T]hat a falsification has not been done conclusively does not mean that it has not been done correctly" (p. 24). The important thing about basic statements, Miller (1982) pointed out, is that they should be true. If there is doubt about a basic statement, it is rational to test it. It is not enough simply to doubt, because doubt is not the same thing as criticism.

Gambling with Nature

The philosophical underpinnings of the demand for severity in testing hypotheses has been discussed and codified by Mayo (1996). "What are needed," she wrote (Mayo, 1996), are arguments that *H* is correct, that experimental outcomes will very frequently be in accordance with what *H* predicts–that *H* will very frequently succeed...We obtain such experimental knowledge by making use of probabilities–not of hypotheses but probabilistic characteristics of experimental testing methods (e.g., their reliability or severity)" (p. 122).

Mayo (1996) explained, "The control of error probabilities has fundamental uses in learning contexts. The link between controlling error probabilities and experimental learning comes by way of the link between error probabilities and severity. The ability to provide methods whose actual error probabilities will be close to those specified by a formal statistical model, I believe, is the key to achieving experimental knowledge" (p. 411).

Mayo seemed to concur with Peirce in this, including Peirce's focus on verification.

Yet, as we have seen, inductive support is not possible. Miller (1982) described the task of empirical science as separating as best it can true statements about the world from false ones, and to retain the true ones. The mission, of course, is to classify, and not certify, truths. Scientific conjectures are "hopelessly fallible, hopelessly improbable, hopelessly unlikely to be true" (Miller, 1982, p. 20). And yet, the conjectural nature of our hypotheses makes them ready to be shown to be wrong. In so doing, we must strictly control the rate at which we make errors in order to ensure a desired level of severity. This imposition of severe testing is a methodological one (Miller, 1982), and it is consistent with both Peirce's philosophical views and with Neyman's (1957) philosophy of inductive behavior.

Neyman (1957) wrote that the concluding phase of scientific research, often labeled inductive reasoning, involves mental processes that are very different from those involved in proving a theorem. Instead of inductive reasoning, which may be considered a misnomer, Neyman preferred the phrase inductive behavior. Neyman pointed out that theories are models of natural phenomena, that is (Neyman, 1957, p. 8)

A model is a set of invented assumptions regarding invented entities such that, if one treats these invented entities as representations of appropriate elements of the phenomena studied, the consequences of the hypotheses constituting the model are expected to agree with observations.

In describing the concluding phase, which he pointed out was frequently described as induction, he felt that the constituent processes were of three types (Neyman, 1957, p. 10). First, the visualization of several possible sets of hypotheses relevant to the phenomenon, second deductions from these sets of hypotheses, and third an "act of will or a decision to take a particular action, perhaps to assume a particular attitude towards the various sets of hypotheses." We need to specify in advance the desired properties of our decision procedure and try to determine the decision rule that has these properties. Given that the hypothesized model is adequate, probability calculations are used to "tell us how frequently the given rule will prescribe any of the actions contemplated" (Neyman, 1957, p. 18). The mental processes involved in the third step, according to Neyman, amount to taking a calculated risk.

Levi (1980) commented on the connection between Peirce's approach to induction and the Neyman and Pearson theory of hypothesis testing: "Peirce's inductions are inferences according to rules specified in advance of drawing the inferences where the properties of the rules which make the inferences good ones concern the probability of success in using the rules. These are features of the rules which followers of the Neyman-Pearson approach to confidence interval estimation would insist on" (p. 138). Peirce's call for predesignation is echoed in Pearson's (1936) insight that "to base the choice of the test of a statistical hypothesis upon an inspection of the observations is a dangerous practice; a study of the configuration of a sample is almost certain to reveal some feature, or features, which are exceptional if the hypothesis is true" (p. 317). Mayo (1993), in drawing out the common philosophical underpinnings of the Peirce and the Neyman-Pearson methodologies, noted that Birnbaum and Armitage showed that violating predesignation permits tests which can be wrong with extremely high probability.

It may be illustrative to view the appropriate use of statistical methods in the course of taking Neyman's calculated risk as a system to use, similar say to a system for playing blackjack, while "gambling with truth" (Levi, 1967) in what Milnor (1954) called "games against nature." In a sense, probability theory is returned to its roots. If the game against nature is to be played, it seems only rational to adopt a system that is known to yield a particular advantageous probability of winning.

In blackjack, even the best systems yield an overall probability of winning of 0.51 or so (Epstein, 1977), so a player must follow a system rigorously or the chances of winning will be reduced, if not reversed. The system is not totally rigid, in that each decision is based on the available information at the time the decision is to be made, but this adaptive decision-making scheme is figured into the overall winning probability, which is known in advance. The player must be steeled against following

intuition or building up superstitions. If a high card is needed, and if the cards so far observed indicate that there is a sufficient proportion of high cards left in the deck to require the player to request a card, the decision should not be influenced by having seen the previous three players receive high cards; nor by the memory that taking a card in a previous similar circumstance led to a losing hand; nor by the feeling that the queen of diamonds is an unlucky card.

Analogously, if prior theoretical or empirical information led on the basis of superior power in a three-group design to the choice of Fisher's (1935) Least Significant Difference (LSD) method of planned comparisons, then that must be the procedure that is carried out. There will be losing hands, experiments in which the Holm procedure would have found significant results that LSD missed. But unless the background information that led to the choice of LSD is substantially changed, the researcher must be comforted by the knowledge that the gambling system that is being employed will in the long run yield errors at the low prespecified rate. On the other hand, if the researcher chooses between LSD and Holm, say, only after the data are seen, the control of error rates is lost. As Miller wrote (1994), "Of course, we can be less zealous, and criticize more mildly. That will not disqualify the proposals that would survive harsher criticism...But it will inevitably compromise the rationality of the decision-making process" (p. 43).

Other well-known examples of the price paid in violating predesignation involve the choice of a one-tailed test (and direction) after the results are known or the choice of a significant covariate for use in an analysis of covariance in the same data set, both of which would increase the Type I error rate. Freedman (1983) similarly found that screening for potential predictors in regression analysis before a final model is fit and tested results in inflated Type I error rates (this result applies to the previous example of covariate choice), and Zimmerman (1996) showed that choosing between Student's t test and the Welch (1947) test on the basis of a test of homogeneity of variance results in a two-stage procedure whose Type I error rates are inflated. Similar problems would arise when the choice between analysis of covariance and analysis of variance is made on the basis of results of tests for baseline differences, (This is especially peculiar when the baseline test is performed even when *random assignment was used*, because in that case the only conclusion to draw is that the randomization was not successful. Should we redo the randomization until we like the results?) or when the choice between the *t* test and a particular form of nonparametric test is made on the basis of the skewness and kurtosis of the dependent variable in the current sample.

The reason that error rates are changed as a result of any similar two-stage procedure is that the first stage test incurs its own errors, which are then compounded in the second stage. Consider Zimmerman's results. If the population variances are equal and the other assumptions of the t test hold, then Student's t test is optimal in holding its Type I error rate and yielding desired power. But the error characteristics of the t test are based on all possible samplings, some of which will yield two samples with apparently different variances. If, in this case, the preliminary test commits a Type I error of its own, the Welch test used at the second stage has lower power than it should, and these cases are also removed from the sampling distribution of the t test. The t is left to operate only on samples whose variances are too close. Conversely, if the population variances are unequal, a Type II error at the first stage results in the use of the *t* test when it is inappropriate, yielding an inflation of the Type I error rate of the method.

Mayo (1993) also pointed out that Pearson, whom she said shied away from Neyman's notion of inductive behavior, 'specifically denied that the tests are to be used as automatic routines for testing claims" (p. 171). Indeed, in this regard, Neyman (1957) criticized Fisher's significance testing approach of having an automatic character in apparently always selecting a one per cent *p*-value as the cutoff for significance, concluding, "There are weighty arguments against this automatism. In fact, it appears desirable to determine the level of significance in accordance with quite a few circumstances that vary from one particular problem to the next" (p. 12). These would

include a consideration of the severity of the errors, both Type I and Type II. Rosnow and Rosenthal (1989, p. 1277) may have been right in this connection when they wrote, "Surely, God loves the .06 nearly as much as the .05", but once they have decided in advance of experiment on a value that would not be too displeasing to the statistical deity, they must ensure that the methods they choose control the error rate at this level.

Mayo (1993) observed that predesignation is only called for when violating predesignation would conflict with the goal of controlling the error probabilities. One example of the use of changing error rates mid-experiment that does not affect the overall properties of the test of a theoretical hypothesis is seen in the context of multiple comparisons. A family is defined as the set of comparisons, the significance of any one of which would lead to the conclusion that the theory has been discorroborated.

Any contrast whose significance does not impinge on the truth of the theory under test is not part of the family. Darlington's (1990) notion of conceptual dependence, to be distinguished from statistical dependence, among contrasts that constitute a family may be helpful in deciding whether or not contrasts belong to a family. Because methodology must be committed to controlling the rate at which the theory is falsely rejected, all legitimate multiple comparison procedures do so successfully, usually through the use of the Dunn-Bonferroni or the improved Dunn-Sidak procedure. (The Bonferroni inequality is due to Boole. Cox, 1977, suggested a sequential adjustment of alpha like the one that is due to Holm, 1979. He gave credit for the suggestion to test the most significant comparison at a Dunn-protected alpha to Tippett in 1931, whereas O'Neill and Wetherill, 1971, call the Dunn-Bonferroni procedure Fisher's Significant Difference method, attributed to Fisher, 1935. For some reason, Dunn's name is too often not included in references to these methods of error rate control.)

Control at the familywise level assures that the probability that one or more of the comparisons is falsely rejected is at most the desired alpha. Because the false rejection of one or more of the comparisons would lead to the false discorroboration of the theory under test, it is this error rate that must be controlled. Any of the sequentially rejective testing procedures, such as those of Holm (1979) or Shaffer (1986), adjusts the Type I error rate assigned to the test of particular comparisons as a function of the results that have been obtained prior to the test of the particular comparisons. This is legitimate, however, because the rate of false discorroboration of the theory is still controlled at the desired level, which itself must be predesignated.

Recently, some interest has been shown in the false discovery rate (FDR) multiple comparison procedure of Benjamini and Hochberg (1995). The FDR is the expected proportion of rejections that are false. Shaffer (1995) suggested that a common misconception, that alpha refers to the proportion of the rejected hypotheses that have been falsely rejected, may have been the reason for the interest in defining and controlling FDR. Benjamini and Hochberg (1995) concluded that familywise (FWE) control is important "when a conclusion from the various individual inferences is likely to be erroneous when at least one of them is" (Benjamini & Hochberg, 1995, p. 290), as, of course, did Peirce and Neyman and Pearson. Benjamini and Hochberg (1995) showed that when all of the hypotheses associated with the multiple comparisons are true, and so the omnibus null hypothesis is true, FDR is equal to FWE, and so in this crucial circumstance, the two procedures are equally viable.

There are other circumstances, Benjamini and Hochberg (1995) felt, in which the less stringent control of FDR is acceptable, such as in exploratory analyses, especially screening problems in which it is desired to obtain as many potential discoveries as possible, but at a controlled rate so as not overly to burden the later confirmatory stage. When considering the different approaches that may be used in exploratory as compared with confirmatory analyses, it is helpful to place the analyses in the context of Peirce's abductions and inductions or of Popper's conjectures and refutations. Because there is an open admission policy toward hypotheses, there is no need for any conjectured relationship to pass a preliminary test, except for

reasons of economy. In the abductive phase, then, any level of alpha can be used that suitably reduces the number of variables later to be tested in an independent study, even values far higher than the conventional five percent level. In the confirmatory stage, however, it is absolutely essential to decide on low and predesignated values of the Type I and Type II error rates, so that the tests are as severe as possible.

Satisficing

In order to test a theory in isolation, instead of as a mix of theory, initial conditions, and auxiliary theories, one must specify in advance of experiment that aspect of the theory that is under test and to assign the remainder, including theories of measurement, to unproblematic background knowledge. To deal with the theory-ladenness of observations, one must remember that the observations are interpreted in terms of theories, including the theory under test. In order to subject the theory to a severe test, we must specify in advance of the experiment what the potential falsifiers of the theory will be, what observational outcomes of the experiment will cause us to regard the theory as falsified.

One of Peirce's rules regarding induction, the inferential method by which hypotheses are tested, is that of predesignation: the property for which a sample is proposed must be specified before sampling, for otherwise "it will always be possible to find some character, however obscure, in which the instances sampled agree, and whether the same proportion of the entire class...has the property will be simply a matter of accident" (Buchler, 1939, p. 246). Indeed, without predesignation, "the induction can serve only to suggest a question, and ought not to create any belief" (Peirce, 1883, p.436).

Peirce (1958) wrote, "The essential thing is that it shall not be known beforehand, otherwise than through conviction of the truth of the hypothesis, how these experiments will turn out" (p. 58). In this regard, Berkson's (1938) observation is pertinent, that if "the result of the...test is known, it is no test at all!" (p. 537). But as discussed previously, it *is* known that the probablility associated with a universal generalization is zero. Recall that in Peirce's view, no theory is true, that Ritchie showed that the probability of any inductive generalization is zero, and that Carnap found that the degree of confirmation of all genuinely universal propositions was zero. Additionally, Peirce claimed that laws of Nature, expressed as simple formulae relating physical phenomena, "are not usually, if ever, exactly true" (Peirce, 1878, p. 334), and finally, Lakatos (1978) opined "that precise particular numerical predictions would have zero measure" (p. 139). Such views are not only expressed by philosophers, and the transfer to statisticians' views concerning the null hypothesis is fairly straightforward. For example, Kempthorne (1976) similarly offered that "A potentially mystifying aspect of this process is that no one, I think, really believes in the possibility of sharp null hypotheses–that two means are absolutely equal in noisy sciences" (p. 772), and Anscombe (1956) wrote that "no one expects any scientific theory to be complete and exact (p. 25).

There are those who defend the possibility of the truth of the point null hypothesis. For instance, Frick (1995) offered as an example of a true point null hypothesis one involved in testing for evidence of extrasensory perception (ESP), and Wainer (1999) considered the case of measuring the speed of light in two reference frames, wherein it is hypothesized that light speed is the same in both experiments. Of note is the fact that the claimed truth of both of these point null hypotheses is based on the assumption of truth of the theories under test, dubious at best given the fallible nature of all knowledge. In terms of the test involving the speed of light, it has been conjectured (Webb et. al., 2001) that certain physical constants such as the speed of light, Planck's constant, and the charge of the electron have been decreasing with time. And if the speed of light were decreasing, then the hypothesis that the two experiments would yield the same value would be false, unless the experiments were conducted simultaneously, again difficult according to the special theory of relativity. The point to be emphasized is that the falseness of point null hypotheses is consistent with the fallibility of theories.

In the case of Frick's ESP example, assume for the sake of argument that ESP is

indeed not possible. In order to test this hypothesis, a person is assigned to guess pictures drawn on a set of cards that are held up in a random order, and the actual content of the card and the guess are recorded. It would be expected that if the cards are selected and the guesses are made at random, there would be zero correlation between them. Unfortunately, neither the guesses nor the card selection are truly random. Diaconis and Mosteller (1989) pointed out that "subjects guess in a notoriously nonrandom manner' (p. 856). Similarly, the order of card selection would be made on the basis of a random device, say a pseudo-random number generator, whose properties are excellent but not perfect. Indeed, MacLaren (1992) showed that the usable length of a pseudorandom sequence was the two-thirds power of its period, after which the uniformity of the sequence no longer conforms to that of a true random sequence. Therefore, the nonrandom sequences of guesses and cards selected will evidence a nonzero correlation. In any experiment, not only must the theory under consideration be true in all respects, but all other aspects of the conditions of experiment would have to be perfectly controlled in order that the value specified in the point null hypothesis be true. This is not at all likely to occur.

This is not to say that it can not happen. The complement to Peirce's previously cited insight that there is a difference between certainty and a probability of unity is that an event whose probability is zero is not impossible. Consider being handed a lottery ticket. If there are a finite number of possible winners, then you have a finite probability of holding the ticket with the winning number. But if the population of possible winning numbers is truly infinite, then your probability of winning is zero, despite your having an actual ticket in your hand. Analogously, although it is not impossible that the numerical value specified in a point null hypothesis is equal to the population parameter, the probability that they are equal for an infinite population is zero.

As a possible solution to the dilemma posed by false point null hypotheses, Lakatos (1978) suggested, "One could...argue...that confirmation theory should be further restricted to predictions within some finite interval of error (p. 139). Similarly, Anscombe (1956) concluded that "we expect some discrepancy between the deduced theoretical hypothesis and our observations. We wish to know if the agreement of observation with hypothesis is *good enough* (p. 25). This notion of specifying a range within which an effect is essentially zero corresponds to Simon's (1957) principle of satisficing and Serlin and Lapsley's (1985) good-enough principle. As an example of the application of the satisficing principle, consider the eclipse experiment in which Einstein's General Theory of Relativity was found to have greater predictive power than Newton's theory (Dyson et. al., 1920). The conclusion that light seemed to be bent by a gravitational object according to Einstein's theory was acclaimed by Thomson (1919) as the most important result obtained in connection with the theory of gravitation since Newton's day" (p. 389). Yet the average of the four widely differing experimental values was off by 10% from theoretical prediction. When asked about the discrepancy, Einstein said that for the expert, this thing is not particularly important.

It is felt that our best theories are close to the truth, that is, that they evidence verisimilitude, and perhaps that over time our theories become closer approximations to the truth. It is necessary to shift our focus to providing a method that allows the conclusion that the theory under test is better than the old one, or that a single prediction is closer to the truth, rather than simply that the difference is nonzero or that the prediction is in error. We could, of course, be wrong. But the emphasis here is on drawing a conclusion concerning the magnitude of an effect. As Anscombe (1956) wrote in this regard, "When testing a theoretical hypothesis, should we not in any case begin by treating the problem as one of estimation, by estimating the magnitude of departure from the theoretical hypothesis" (p. 25). Often, the hypothesis test and the estimation of magnitude are considered separate parts of the analysis. For example, Yates (1948) noted, "The first point that struck the practical man was that experiments in general performed two different functions, one being to test the significance of a certain hypothesis, and the other to estimate the magnitude of the deviation from that hypothesis

if, in fact, it was found to be, or was suspected of being, untrue" (p. 204).

One reason for this apparent disconnect between hypothesis testing and estimation by confidence interval is that the traditional point null hypothesis only allows the conclusion that the parameter is not exactly as specified, whereas the essential information to be obtained in an experiment regards whether the parameter is outside of the good-enough region. Unfortunately, the classical Neyman-Pearson confidence interval can not answer this question well. In the traditional case, it is posited that the test statistic has a certain distribution, given that the parameter is equal to a specific value, and the inversion of this distribution yields the confidence interval for the parameter, given the observed test statistic. But the results of the hypothesis test can be significant, indicating a nonzero effect, without the confidence interval indicating that the magnitude of the effect is important.

Of course, the logic underpinning the standard confidence interval is solid. We can legitimately reason that if the population mean equals a particular value, then given the data, the confidence interval can be derived using the solid statistical principles offered by Neyman and Pearson. The logic is impeccable. But because the value specified in a point null hypothesis has zero probability of being correct, Descartes might have said, "I don't doubt the validity of your inference, only the premise."

Equally troubling is the finding by Meeks and D'Agostino (1983) that the coverage probability of the classical confidence interval is liberal if one only constructs the confidence interval after rejection of the point null hypothesis. Instead, if the confidence interval is derived from the inversion of the distribution of the test statistic that would be used to test a range null hypothesis, the interval answers the question of interest regarding whether the magnitude of the effect is large enough, there is a nonzero probability that the range specified in the null hypothesis covers the limit to the population range, and the results of the confidence interval and hypothesis test are consistent. Hodges and Lehmann (1954) and Serlin and Lapsley (1985, 1993) provided tests of range null hypotheses that allow the conclusion that an effect is large enough. An example of the use of a range null hypothesis test to show large effects was provided by MacCallum, Browne, and Sugawara (1996) in the context of covariance structure modeling. Examples of the use of confidence intervals that provide good-enough information are given in Steiger and Fouladi (1997), Cumming and Finch (2001), Fidler and Thompson (2001), and Smithson (2001).

In addition, range null hypotheses (and confidence intervals) can be used to examine theories that predict effects of at least a certain magnitude by allowing the discorroborating conclusion that the effect is smaller than that demanded by the theory. The bioequivalence literature introduced many tests that allow the conclusion that an effect is small, as did Serlin and Lapsley (1985, 1993), Rogers, Howard, and Vessey (1993), and Seaman and Serlin (1998). Serlin (2000) showed how such a test could be used in a Monte Carlo study to establish that a statistical procedure satisfies specified criteria for robustness. As previously indicated for the general case, in using any of these procedures, the criterion for a large enough effect or an effect that is small enough to discorroborate the theory must be predesignated.

Implications for future research

In his book on games of chance, according to David (1962), Cardano lamented that the facts of probability that he discovered contribute to mathematical understanding but not to the gambler. It has been shown, however, that quite to the contrary, the theory of probability is essential to a rational scientific methodology in the game against nature. Point null hypotheses, like universal theories, are quite probably false, as are the assumptions underlying statistical tests. As Cox (1958) wrote, "Assumptions that we make, such as those concerning the form of the populations sampled, are always untrue" (p. 369). It is essential, then, that we be able to examine the verisimilitude of theories through the application of severe range null hypothesis tests whose assumptions are themselves subjected to serious scrutiny. The *Journal of Modern Applied Statistical Methods* is particularly well-placed to advance statistical methodology in this regard.

In order to conduct a severe test of a hypothesis, the Type I error rate of the statistical procedure must be held as close as possible to its predesignated size, and the power of the test must not fall far from its specified level, regardless of the nature of the populations sampled. To this end, robust procedures for testing range null hypotheses have to be developed and investigated. The most difficult problem to be addressed likely will involve finding a means to incorporate the hypothesized good-enough range, expressed in actual or standardized units of the raw scale, into the distribution-free procedure.

For example, in a one-sample test that a theoretical prediction is no more than 0.2 standard deviations from the true value, the satisficing range must be introduced in both the hypothesis to be tested and the sampling distribution of the test statistic. The satisficing limit of 0.2 standard deviations must be expressed in terms of the population median for the range null hypothesis addressed by the signed-rank Wilcoxon test, and the null range must also be incorporated into the sampling distribution of the signed-rank statistic. Similar accommodations must be made in a multiple-sample, multiple-predictor, and/or multiple dependent variable test in which the null range is specified in terms of a measure of association, such as R-squared, or in terms of a function of eigenvalues or the Mahalanobis distance.For instance, if the range null hypothesis is stated in terms of the squared multiple correlation coefficient between a set of predictors and a dependent variable, what are the corresponding parameters and sampling distribution of the sample statistic in a rank regression test of the appropriate range null hypothesis?

Regardless of the nature of the hypotheses and tests, the assumptions underlying the procedures must be taken into account. In the one-sample case, asymmetric pre- and post-tests with unequal variances will yield asymmetric difference scores, which would violate the assumptions underlying the matched-pair Wilcoxon test, as would having a single asymmetric dependent variable. As with the matched-pair Wilcoxon test, the properties of the adjusted Mann-Whitney test of Fligner and Policello (1981) and the modified Kruskal-Wallis test of Rust and Fligner (1984), which accommodate unequal variances in multiple-group tests of location, are affected by asymmetry. Although much work has been done in this regard, the properties of tests of symmetry seem to depend on other properties of the distribution, such as kurtosis (Antille, Kersting, & Zucchini, 1982; Fan & Gencay, 1995; Brizzi, 2002), and so more work in this area is needed. In addition, differing variances and covariances of sets of difference scores in a repeated measures design violate the assumptions of the Friedman test and other competitors (Harwell & Serlin, 1994). The multiple group, multiple measure design would analogously require nonparametric tests of sphericity and homogeneity of covariance matrices, as would the test of identity of regression lines and the test of parallelism that is used to examine hypotheses concerning moderating variables.

Most importantly, the need for range null hypothesis tests applies both to the test of theory and to the tests of assumptions. That is, the requirement of satisficing applies at all levels of the scientific endeavor. Because theories are improbable, a good-enough region must be determined in advance of experiment, so that potential falsifiers can be specified. This, in turn, requires that a range null hypothesis be tested, in order to determine if a discorroborating outcome has occurred. And the test can only be considered severe if the error probabilities are held within an acceptable range of the predesignated levels, according to a criterion of robustness.

When examining whether or not the assumptions underlying a statistical procedure are satisfied, the hypothesis to be tested concerning the assumptions must specify that the statistical model that is conjectured to apply to the data is a good enough fit, that is, that the assumptions underlying the statistical test of a substantive theory are met well enough that the statistical test itself meets its criterion of robustness. This means that a good-enough region must be specified in a range null hypothesis of the test of the validity of the assumptions underlying the statistical test of the substantive theory, and robust tests of these range null hypotheses concerning assumptions

need to be developed. To this end, Monte Carlo studies of the robustness of procedures must provide response surfaces reflecting the Type I error rate and power as a function of the inexact agreement of model and data. Pearson and Please (1975), for example, present the Type I error rates for the one- and two-tailed, one- and two-sample *t* tests and tests of variances in a series of graphs for varying kurtosis at specific values of skewness. A researcher could determine limits to the skewness and kurtosis that lead to the two-sample *t* test, say, meeting a criterion for robustness; then these limits, in turn, would be implemented in range null hypotheses in a pilot study to determine if the skewness and kurtosis of the distribution of the population from which the proposed sample is to be drawn adequately meet the requirements for robustness of the *t* test.

Conclusion

Attempts to attain knowledge as certified true belief have failed to circumvent Hume's injunction against induction. Unfortunately, Hume also showed that the search for probable knowledge, that which Locke called opinion or belief, also depended on an inductive principle. Instead, theories must be viewed as unprovable, improbable, and undisprovable (Lakatos, 1970) because, in addition to Hume's criticism of justificationism, Peirce among others showed that the empirical basis is fallible. Importantly, though, as Whewell advocated, the method of conjectures and refutations is untouched by Hume's insights.

The implication for statistical methodology is that the requisite severity of testing is achieved through the use of robust procedures, whose assumptions have not been shown to be substantially violated, to test predesignated range null hypotheses. Nonparametric range null hypothesis tests need to be developed to examine whether or not effect sizes or measures of association, as well as distributional assumptions underlying the tests themselves, meet satisficing criteria.

References

Agassi, J. (1975). Subjectivism: From infantile disease to chronic illness. *Synthese*, *30*, 3-14.

Agassi, J. (1983). Theoretical bias in evidence: A historical sketch. *Philosophica*, *31*, 7-24.

Anscombe, F. J. (1956). Discussion on Dr. David's and Dr. Johnson's paper. *Journal of the Royal Statistical Society*, Series B, *18*, 24-27.

Antille, A., Kersting, G., & Zucchini, W. (1982). Testing symmetry. *Journal of the American Statistical Association*, *77*, 639-646.

Bartley, W. W. III. (1984). *The retreat to commitment*. LaSalle, Illinois: Open Court.

Bell E. T. (1937). *Men of mathematics*. New York: Simon & Schuster.

Bellhouse, D. (1993). The role of roguery in the history of probability. *Statistical Science*, *8*, 410-420.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society,* Series B, *57*, 289-300.

Bennett, J. H. (Ed.) (1990). *Statistical inference and analysis*. Oxford: Clarendon Press.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33*, 526-542.

Bradley, J. (1971). *Mach's philosophy of science*. New York: Oxford University Press.

Brizzi, M. (2002). Testing symmetry by an easy-to-calculate statistic based on letter values. *Developments in Statistics*, *17*, 63-74.

Buchler, J. (1939). *Charles Peirce's empiricism*. New York: Harcourt, Brace and Company.

Burtt, E. A. (1924). *The metaphysical foundations of modern physical science*. London: Routledge.

Clark, G. H. (1957). *Thales to Dewey*. Boston: Houghton-Mifflin.

Cohen, R. S. (1970). Ernst Mach: Physics, perception and the philosophy of science. In R. S. Cohen & R. J. Seeger (Eds.), *Boston studies in the philosophy of science*, *Volume VI*, pp. 126-164. Dordrecht-Holland: Reidel.

Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics*, *29*, 357-372.

Cox, D. R. (1977). The role of significance tests. *Scandinavian Journal of Statistics*, *4*, 49-70.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532-74.

Darlington, R. B. (1990). *Regression and Linear Models*. New York: McGraw Hill.

David, F. N. (1962). *Games, Gods & Gambling*. New York: Hafner.

Descartes, R. (1642/1927). Meditations. In R. M. Eaton (Ed.), *Descarte selections*. New York: Charles Scribner's Sons.

Diaconis, P., & Mosteller, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association*, *84*, 853-861.

Dyson, F. W., Eddington, A. S., & Davidson, C. (1920). A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society of London*, *220*, 291-333.

Epstein, R. A. (1977). *The theory of gambling and statistical logic*. New York: Academic Press.

Fan, Y., & Gencay, R. (1995). A consistent nonparametric test of symmetry in linear regression models. *Journal of the American Statistical Association*, *90*, 551-557.

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement*,*61*, 575-604.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. New York: Macmillan.

Fligner, M.A., & Policello, G. E. III (1981). Robust Rank Procedures for the Behrens–Fisher Problem. *Journal of the American Statistical Association*, *76*, 162–168.

Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician*, *37*, 152-155.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition*, *23*, 132-138.

Fuller, B. A. G. (1938). *A history of philosophy*. New York: Henry Holt and Company.

Garber, D. (1995). Apples, oranges, and the role of Gassendi's atomism in seventeenth-century science. *Perspectives on Science*, *3*, 425-428.

Gillies, D. (1993). *Philosophy of science in the twentieth century*. Oxford: Blackwell.

Harris, J. F. (1992). *Against relativism*. LaSalle, IL: Open Court.

Harwell, M. R., & Serlin, R. C. (1994). A Monte Carlo study of the Friedman test and some competitors in the single factor, repeated measures design with unequal covariances. *Computational Statistics & Data Analysis*, *17*, 35-49.

Hodges, J., & Lehmann, E. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society (B)*, *16*, 261-268.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65-70.

Jeffreys, H. (1939). *Theory of probability*. London: Oxford University Press.

Joy, L. S. (1995). Rationality among the friends of truth: The Gassendi-Descartes controversy. *Perspectives on Science*, *3*, 429-449.

Kadane, J. B. (1976). For what use are tests of hypotheses and tests of significance, introduction. *Communications in Statistics* (A), *5*, 735-736.

Kempthorne, O. (1976). For what use are tests of significance and tests of hypothesis. *Communications in Statistics,* Part A, *5*, 763-777.

Kiernan, J. F. (2001). Points on the path to probability. *The Mathematics Teacher*, *94*, 180-183.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In Imre Lakatos & Alan Musgrave (Eds.), *Criticism and the growth of knowledge*, 91-196. Cambridge: Cambridge University Press.

Lakatos, I. (1978). Newton's effect on scientific standards, in J. Worrall & G. Currie (Eds.): *The methodology of scientific research programmes*, 193-222. Cambridge: Cambridge University Press.

Levertov, D. (1961). Matins. *The Jacob's ladder.* New York: New Directions.

Levi, I. (1967). *Gambling with truth*. New York: Knopf.

Levi, I. (1980). Induction as self correcting according to Peirce. In D. H. Mellor, *Science, belief, and behavior: Essays in honor of R. B. Braithwaite*, 127-140. Cambridge: Cambridge University Press.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130-149.

MacLaren, N. M. (1992). *Journal of Statistical Computation and Simulation*, *42*, 47-54.

Mayo, D. G. (1993). The test of experiment: C. S. Peirce and E. S. Pearson. In E. C. Moore (Ed.), *Charles S. Peirce and the philosophy of science*, 161-174.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Medawar, P. (1974). Hypothesis and imagination. In P. A. Schilpp (Ed.), 274-291. LaSalle, IL: Open Court.

Meeks, S. L. & D'Agostino, R. B. (1983). A note on the use of confidence limits following rejection of a null hypothesis. *The American Statistician*, *37*, 134-136.

Miller, D. (1980). Can science do without induction? In L. J. Cohen & M. Hesse (Eds.), *Applications of inductive logic*, 109-129.

Miller, D. (1982). Conjectural knowledge: Popper's solution of the problem of induction. In P. Levinson (Ed.), *In pursuit of truth*, 17-49. New Jersey: Humanities Press.

Miller, D. (1994). *Critical rationalism*. Chicago: Open Court.

Miller, D. (2002). Induction: A problem solved. In J. M. Böhm, H. Holweg, & C. Hoock: *Karl Popper's kritischer rationalismus heute*, 81-106. Tübingen: Mohr Siebeck.

Milnor, J. (1954). Games against nature. In R. M. Thrall, C. H. Coombs, & R. L. Davis (Eds.), *Decision processes*, 49-59.

Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research*, *22*, 267-305.

Musgrave, A. (1993). Popper on induction. *Philosophy of the Social Sciences*, *23*, 516-527.

Neyman, J. (1957). "Inductive behavior" as a basic concept of philosophy of science. *International review of statistics*, *25*, 7-22.

Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society*, A, *231*, 289-337.

O'Neill, R., & Wetherill, G. B. (1971). The present state of multiple comparison methods. *Journal of the Royal Statistical Society,* Series B, *33*, 218-250.

Owen, D. (1993). Locke on reason, probable reasoning, and opinion. *The Locke Newsletter*, *24*, 35-79.

Pearson, E. S., & Sekar, C. C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, *28*, 308-320.

Pearson, E. S., & Please, N. W. (1975). Relation Between the Shape of Population Distribution and the Robustness of Four Simple Test Statistics. *Biometrika*, *62*, 223-241.

Peirce, C. S. (1868). Some consequences of four incapacities claimed for man. *Journal of Speculative Philosophy*, 2, 140-157.

Peirce, C. S. (1878). Deduction, induction, and hypothesis. In C. J. W. Kloesel (Ed.), *Writings of Charles S. Peirce*, *3*, 323-338. Bloomington, Indiana: Indiana University Press.

Peirce, C. S. (1883). A theory of probable inference. In C. J. W. Kloesel (Ed.), *Writings of Charles S. Peirce*, *4*, 408-450. Bloomington, Indiana: Indiana University Press.

Peirce, C. S. (1958). *Collected papers of Charles Sanders Peirce*. Edited by A. W. Burks. Vol. VII.: *Science and Philosophy*. Cambridge: Harvard University Press.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Popper, K. (1962). *Conjectures and refutations*. New York: Basic Books.

Popper, K. (1985). *Realism and the aim of science*. From W. W. Bartley III (Ed.), *Postscript to the logic of scientific discovery*. London: Routledge.

Quine, W. V. O. (1951). Two dogmas of empiricism. Reprinted in *From a logical point of view*. (2nd rev. ed.). Harper Torchbooks, 20-46.

Radnitzky, G. (1982). Popper as a turning point in the philosophy of science: Beyond foundationalism and relativism. In P. Levinson (Ed.), *In pursuit of truth*, 64-80. Sussex: Harvester Press.

Reichenbach, H. (1951). *The rise of scientific philosophy*. Berkeley: University of California Press.

Ritchie, A. D. (1926). Induction and probability. *Mind*, *35*, 301-318.

Rogers, J., Howard, K., and Vessey, J. (1993). Using Significance tests to evaluate equivalence between experimental groups. *Psychological Bulletin*, *11*, 553-565.

Rosnow, R, L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276-1284.

Russell, B. (1945). *A history of western philosophy*. New York: Simon and Schuster.

Rust, Steven W., & Fligner, Michael A. (1984), A modification of the Kruskal-Wallis statistic for the generalized Behrens-Fisher problem. *Communications in Statistics,* Part A -- Theory and Methods, *13*, 2013-2027.

Salmon (1966). The foundations of scientific inference, in R. G. Colodny (Ed.): *Mind and Cosmos*, 135-275. Pittsburgh: University of Pittsburgh Press.

Seaman, MA, & Serlin, RC (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, *3*, 403-411.

Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods*, 5, 230-240.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*, 73-83.

Serlin, R. C. & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences*, 199-228. Hillsdale, N. J.: Erlbaum.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, *81* , 826-831.

Shaffer, J. P. (1995). Multiple hypothesis testing: A review. *Annual Review of Psychology*, 46, 561-584.

Simon, H. A. (1957). *Models of man, social and rational*. New York: Wiley.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61,* 605-32.

Steiger, H. H., & Fouladi, R. T. (1996). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, and J. H. Steiger, (Eds.), *What if there were no significance tests?,* 221-257. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Stigler, S. M. (1986). *The history of statistics*. Cambridge, Mass.: Harvard University Press.

Suppe, F. (1977). *The structure of scientific theories*. Chicago: The University of Illinois Press.

Sylla, E. D. (1998). The emergence of mathematical probability from the perspective of the Leibniz-Jacob Bernoulli correspondence. *Perspectives on Science*, *6*, 41-76.

Todhunter, I. (1865). *A history of the mathematical theory of probability from the time of Pascal to that of Laplace*. Cambridge: Cambridge University Press.

Thomson, J. J. (1919). Joint Eclipse Meeting of the Royal Society and the Royal Astronomical Society. *The Observatory, 42*, 389-398.

Von Mises, R. (1970). Ernst Mach and the empiricist conception of science. In R. S. Cohen & R. J. Seeger (Eds.), *Boston studies in the philosophy of science*, Volume VI, 245-270. Dordrecht-Holland: Reidel.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*, 212-213.

Walker, H. M. (1929). *Studies in the history of statistical method*. Baltimore: Williams and Wilkins.

Watkins, J. (1978). The Popperian approach to scientific knowledge. In G. Radnitzky & G. Andersson (Eds.), *Progress and rationality in science*, 23-43. Dordrecht, Holland: Reidel.

Webb, K., Murphy, M. T., Flambaum, V. V., Dzuba, V. A., Barrow, J. D., Churchill, C. W. Prochaska, J. X., & Wolfe, A. M. (2001). Further evidence for cosmological evolution of the fine structure constant, *Physical Review Letters*, *87*, 091301-1-091301-4.

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, 29-35.

Wettersten. J. R. (1992). *The roots of critical rationalism*. Atlanta: Rodopi.

Wettersten, J. R. (1993). Rethinking Whewell. *Philosophy of the Social Sciences*, *23*, 481-515.

Wiener, P. P. (1972). *Evolution and the founders of pragmatism*. Philadelphia: University of Pennsylvania Press, Inc.

Yates, F. (1948). Discussion on Mr. Anscombe's paper. *Journal of the Royal Statistical Society,* Series A, *111*, 204-205.

Zimmerman, D. W. (1996). Some properties of preliminary tests of equality of variances in the two-sample location problem. *The Journal of General Psychology*, *123*, 217-231.