RONALD N. GIERE

The Epistemological Roots of Scientific Knowledge

1. Introduction

The objective of this paper is to complete one part of a broad program for resolving all important issues involving probability, inductive logic, and scientific inference generally. I shall begin by sketching the general program and locating the specific problem of interest in the general scheme. After the specific problem has been resolved I will conclude with some further discussion of the general program.

Though metaphilosophy belongs last in the order of discovery, there are good reasons for putting it first in the order of exposition. Let us begin, therefore, with the meta-question: What is the ultimate goal of inquiry into probability and inductive inference? Some have argued that scientific, nondeductive reasoning is necessarily an intuitive piecemeal affair, more art than theory. The most we can accomplish, on this view, is some conceptual clarification and codification of a variety of inductive methods.[1] The goal of the majority, on the other hand, has been to construct a unified, comprehensive account of scientific reasoning that could win the kind of widespread acceptance enjoyed by deductive logic.[2] This is the goal adopted here.

AUTHOR'S NOTE: This work was completed during my tenure as associate research scientist at the Courant Institute of Mathematical Sciences, New York University, and has been supported by the National Science Foundation under grants GS 2525 and GS 28628. A special acknowledgment is due Professor Allan Birnbaum of the Courant Institute both for much helpful criticism and for making it possible to pursue this work in such exciting surroundings.

[1] Among philosophers this viewpoint is generally associated with the "ordinary language" school exemplified by Strawson (1952) and Black (1971). Popperians (1959) also generally reject the idea of a systematic theory of inductive inference, though for very different reasons. Among statisticians, Tukey (1961) is well known for advocating "data analysis" over standard methods of testing and estimation.

[2] This was certainly the aim of both Carnap (1950) and Reichenbach (1949). It is also the aim of later philosophical students of inductive reasoning, e.g.,

The core of any overall account of inductive inference would be a precise *characterization* of the form (or forms) of inductive inferences. The system thus characterized would be an inductive logic in the broad sense that it would be a canon for validating scientific reasoning. But that it would be anything like the inductive logics of Carnap (1950) or Hintikka (1966) is an open question. Indeed, the central question for any account of inductive reasoning seems to me to be whether inductive logic is to be characterized in terms of a *relation*, e.g., degree of belief, between evidence and a hypothesis, or in terms of a *method* for using data to reach a conclusion or to choose a course of action. My understanding of 'inductive logic' is intended to be broad enough to cover both possibilities.

While a precisely characterized inductive logic is the heart of any comprehensive account of scientific reasoning, there are several other crucial components. One is an *interpretation* of the mathematical probability calculus. The problem of interpreting the probability calculus and characterizing an inductive logic are quite distinct, though often closely related. For Reichenbach (1949) or Salmon (1966) the distinction is clear: probabilities are interpreted as limiting relative frequencies while induction is characterized by a rule of simple enumeration. Carnap's system (1950) employs both inductive (logical) probabilities and statistical probabilities, the latter being interpreted as limiting relative frequencies. Even subjective Bayesians like de Finetti (1937) and Savage (1954) face both an interpretation problem and a characterization problem, though here the two are very closely entwined.

Another vital component of a unified view of induction is an account of how inductive reasoning may be used in *practical decision-making*. Some philosophers, e.g., Popper (1962) and his followers, would deny any connection between theoretical scientific reasoning and "engineering." But this seems to me only to show the untenability of the strict Popperian view of science. There has got to be an intelligible connection between knowledge and action; any overall view of scientific inference that does not give an account of this connection is defective. Accounting for the connection, however, is not an easy task. Indeed, the difficulties in constructing a suitable decision theory on other characterizations of induc-

Kyburg (1961), Jeffrey (1965), Salmon (1966), Hintikka (1966), and Levi (1967).

tive inference constitute one of the major attractions of a subjective Bayesian characterization.[3]

Finally there is the philosophical problem of the *justification of induction*. Historically this problem was most forcefully exposed by Hume. He assumed an inductive logic (basically simple enumeration) in which singular or general statements are inferred from singular statements alone. Noting that the connection between premises and conclusions is nondeductive, Hume asked how one can justify drawing the conclusion without making some other nondeductive step for which a similar justification is needed. He concluded that no rational justification is possible. For those wishing to avoid such skepticism, the problem is to characterize a nondeductive form of inference and then to justify using it.

Though it may be couched in logical or psychological terms, Hume's problem is essentially epistemological. The problem is to justify claiming that one knows, or has good reason to believe, the conclusion, given that one knows the evidence. The same sort of problem arises for relational characterizations as well. Thus one who claims that scientific inference has anything to do with acquiring scientific knowledge must provide some kind of justification for using his particular inductive logic to achieve this end. Yet many kinds of justification are possible, even for a single inductive logic.

Granting that one must justify using any particular inductive logic to validate claims to empirical knowledge, there is an even more general meta-problem to be faced. Suppose one has several inductive logics, each with a different kind of epistemological rationale. How is one to choose among these different comprehensive systems? For example, how would one decide between Carnap's system and Reichenbach's system? I call this the problem of determining the *adequacy* of a comprehensive account of inductive reasoning. In what follows it will be essential to distinguish this most general meta-problem from the narrower problem of justifying a particular characterization of inductive reasoning. I take it as a defect of most discussions of "the justification of induction" that these

[3] It is quite clear that for Savage (1954) one of the primary motivations for introducing subjective probabilities is that they automatically reduce the problem of decision-making under uncertainty to that of decision-making under risk. For a good background discussion of these issues see ch. 13 of Luce and Raiffa (1957). A number of philosophers, e.g., Jeffrey (1956), Carnap (1963), and Bar-Hillel (1968), have used the problem of constructing a practical decision theory as the basis for rejecting any characterizations of inductive logic in which hypotheses are 'accepted' or 'rejected'. For a contrary view see Kyburg (1968) and Levi (1967).

two problems are not clearly distinguished. Indeed, many authors tackle the justification problem with little or no attention to the assumed characterization of inductive inference. It is to be expected that such discussions would confuse the problems of justification and adequacy.

How then do we evaluate the adequacy of a unified account of scientific inference? I will discuss this problem further in section 8, but several conclusions are needed here. My view is that judgments of adequacy cannot be reached by *rational* argument without begging the question. Such judgments can, however, be reached through a *causal* process, the appropriate process being scientific research itself. It may be objected that one should no more give in to unformalized intuition, even collective intuition, at the meta-level than at the object level. But not everything can be formalized and rationally justified. It is sufficient for science that there be a well-developed inductive logic for validating scientific results. That the adequacy of the whole system must rest on judgments that are ultimately only caused and not reasoned is unavoidable. Those inclined to lament this state of affairs should consider whether the situation is any different regarding rival systems of *deductive* logic.

It follows from the above that the only judgments of adequacy we can make are evaluations of the *relative* adequacy of unified accounts of scientific inference which are sufficiently well developed to be applied in actual research contexts. This requirement effectively eliminates most philosophical accounts of scientific reasoning as candidates for our inductive logic since these are invariably so schematic and idealized as to be inapplicable even to the simplest cases of actual research. For sufficiently rich inductive logics one must look to the last half century's achievements in theoretical statistics. But here one faces the reverse problem. Statistical methods are widely applicable, and widely applied but there exists no really unified characterization of the logic, only partial characterizations which are not even mutually consistent.[4] Thus before raising any questions of adequacy, one must first reconstruct a suitably rich logic of statistical inference. The same is true of the statis-

[4] Formulating widely held principles of statistical inference in terms of evidential irrelevance, Birnbaum (1969) demonstrates several striking inconsistencies. Birnbaum is troubled by this result and seems pessimistic regarding the prospects for any unified inductive logic. Others, who, like Tukey (1961), refer to themselves as 'data analysts', seem content to pick and choose among apparently inconsistent methods. For an example of this growing eclectic approach, see the recent text by Kempthorne and Folks (1971).

tician's interpretation of probability. There is a lot said in standard treatises about relative frequencies, but few theoretical statisticians hold a strict limiting frequency interpretation a la von Mises (1957) or Reichenbach (1949). Again, some statisticians, following Wald (1950), would define statistics as the theory of practical decision-making using statistical data. But this description is hard to square with the practice of these same statisticians in scientific contexts.[5] Finally, theoretical statisticians, like scientists, have little stomach for questions about ultimate justification. The statistical literature contains little that could be regarded as even an implicit answer to Hume. On the other hand, at least some statisticians have an intuitive appreciation for the problem of judging the relative adequacy of competing systems, though the problem is not precisely stated and there is little meta-discussion of how it might be resolved.[6]

With the background presented above it is possible to give a brief statement of my general program for resolving the fundamental issues concerning scientific inference. It is to develop rival approaches to scientific inference into systematic, unified accounts of scientific reasoning so that they may be compared in typical scientific research contexts. Beyond this one can only encourage theoreticians and practitioners of scientific inference to make the same comparisons and hope that there is a de facto convergence of opinion on the relative adequacy of the rival systems.

Judgments of adequacy require precisely developed and comprehensive accounts of scientific inference which as yet do not really exist. So the immediate task is one of construction, or reconstruction, and development. Now there are several extant prototype inductive logics which can deal with interesting scientific cases. There is, first of all, the Fisherian tradition which includes maximum likelihood point estimation, hypothesis testing, and fiducial inference. Then there is the Neyman-Pearson account of testing and interval estimation. The latter is

[5] See, for example, ch. 1 of Lehman (1959). Neyman must be regarded as the founder of this school. Yet an examination of his scientific work in such fields as astronomy and meteorology reveals little explicit adherence to the decision theoretic point of view. For references to these scientific papers see the bibliography in Neyman (1967).

[6] Savage (1967) is an exception to the rule that statisticians do not take Hume seriously. Also, Birnbaum (1971) has explicitly advocated the use of scientific case studies to help resolve fundamental conflicts regarding the foundations of statistics.

closely related to the Neyman-Wald account of statistical decision theory. The most systematic inductive logic and decision theory is that of subjective Bayesians like Savage, but extensive scientific applications of the Bayesian approach are hard to find. Finally, there are those who, like Tukey, practice 'data analysis' without utilizing any systematic logic.[7]

Of these possibilities, the Neyman-Pearson-Wald tradition seems to me most promising. This is largely because, like Peirce and Reichenbach, I see induction in terms of a method for using data to reach conclusions and not merely in terms of a quantified relation between data and hypotheses. As indicated above, however, there is no single unified account within this tradition. Thus one must reconstruct a plausible inductive logic and decision theory, develop a congenial interpretation of probability, and supply a philosophical justification for using the reconstructed logic to certify scientific knowledge claims. Characterizing the logic of Neyman's approach is the most urgent and difficult of these problems, but I am not prepared to do justice to this task here. I have already done some work on the interpretation of probability. This leaves the justification problem. Of course one cannot discuss justification without presupposing a logic and an interpretation of probability, but enough can be given to provide a basis for raising interesting questions about justification. Moreover, for those not familiar with current statistical theory, and this unfortunately still includes most philosophers — even most philosophers of science — a discussion of the justification problem may provide the better motivation for taking seriously the kind of general program outlined above.

In the following section I will sketch all of scientific inference and argue for the epistemologically fundamental position of statistical inference. I will also give an interpretation of physical probability. The third section contains a brief characterization of a Neyman-like theory of statistical hypothesis testing. This is followed by a discussion of the

[7] Fisher's views are developed in several books (1925, 1935, 1957) and in many papers (1956). Neyman and Pearson's fundamental papers are now reprinted in Neyman and Pearson (1967) and in Neyman (1967). The bible of the subjective Bayesian school is Savage (1954). See also Good (1950, 1965). Less technical presentations, together with considerable bibliography, may be found in Savage (1962) and in Kyburg and Smokler (1964). For a recent exposition by the foremost British Bayesian, see Lindley (1971). The 'data analysts' include Tukey (1961), Birnbaum (1962), and Kempthorne (1971). Finally, one group not mentioned above consists of those, e.g., Hacking (1965) and Edwards (1971), who attempt to build a theory of statistical inference on likelihood ratios alone.

epistemological structure of hypothesis testing. After looking in some detail at a semirealistic example, I examine, in section 6, the prospects for employing a traditional 'foundationist' justification for the indicated logic of testing. Rejecting the foundationist approach, I go on to develop a Peircean 'methodological' justification. In section 8 I return to the relation between justification and adequacy. The conclusion is devoted to some discussion of the metaphors of ultimate justification, i.e., the traditional notion of the "foundations" of knowledge, Quine's "web of belief," and the idea that scientific knowledge may have "roots" even though it has no "foundations."

2. Probability and Physical Systems

The world is a complex physical system. Any aspect of a physical system may be described by a quantity called a *state* of the system. Thus the total state of a system may be represented by a vector in the appropriate *state-space*. A theory of a kind of system includes at least a description of the relevant state-space together with a designation of certain states as physically possible (kinematics) and a function giving the development of the system in time relative to any possible initial state (dynamics). The history of any system is represented by a trajectory in the state-space.[8] Roughly speaking, a system is *deterministic* if truly described by a theory which, for any arbitrarily given initial state, specifies a unique state for any time. A system is *probabilistic* if truly described by a theory which, for any arbitrarily given initial state, specifies a probability distribution over possible states at later times.[9]

In a deterministic system, some states are *physically necessary* relative to a given initial state. I am inclined to hold that this necessity is real, i.e., a feature of the physical system itself and not merely a creature of language. Similarly, I wish to say that a *probabilistic* system has *physical propensities* to evolve into certain states relative to a given initial state. Such propensities are like physical necessities, but somehow weaker. Moreover, on this view there is no direct logical connection between propensities and relative frequencies. The strongest connection is given

[8] This general view of the nature of theories follows van Fraassen (1970), though here more in spirit than in letter.

[9] For an appreciation of the kinds of difficulties that develop when trying to give a precise account of deterministic theories see Earman (1971), Glymour (1971), and van Fraassen (1972).

by an interpretation of the Bernoulli theorem. If a long series of relevantly similar independent systems begin with the same initial state, then the compound system has a very strong propensity to exhibit final states with relative frequencies very near the corresponding propensity. But this relative frequency is not necessary, no matter how long the sequence.[10]

The above is mainly ontology, and the subject here is epistemology. But since I find this ontology congenial and shall assume it for the purposes of exposition at least, it is best to make it explicit. On the other hand, there are probably no necessary connections between my ontology and epistemology, so that the following discussion may easily be reinterpreted by anyone with other, e.g., Humean, ontological commitments.

Turning, then, to epistemology, how does one test theories about physical systems? The traditional view is simply that from our theory of a kind of physical system we deduce characteristics of particular systems whose operations are open to observation. If the observed systems behave as the theory indicates, that is taken as evidence that the overall theory is correct. This characterization of theory testing is extremely general and incomplete. It does not even tell us whether positive results confer probability on the theory or whether they make the theory 'acceptable' without assigning probabilities. On the other hand, its very generality makes the traditional account agreeable to almost everyone, and, fortunately, no more precise account is necessary for present purposes.

It is a commonplace that an experiment never tests a theory alone. Rather, what is tested is the theory together with assumptions about the experimental setup. But putting this point in terms of "auxiliary assumptions" masks the nature of the situation. It is better, as Suppes (1962) seems to suggest, to think of a hierarchy of interacting physical systems described by a corresponding hierarchy of theories, including a theory of the experimental apparatus. Making these theories "mesh" appropriately raises some nice logical and methodological problems, but these are of no special concern here. The crucial point is that the lowest level system is always a *probabilistic* system, i.e., a data generator. The epis-

[10] I have discussed a "single case" propensity interpretation of probability in Giere (1973a), (1975). The most extensive development of any propensity theory to date is due to Mellor (1971). See also my review, Giere (1973c).

temologically relevant hypotheses concerning the experimental setup are therefore *statistical* hypotheses. This makes the testing of statistical hypotheses epistemologically prior to the testing of general theories. One may thus consider the justification of a method of statistical inference as basic to the justification of any complete inductive logic.

The claim that experimental setups are always probabilistic systems need not be understood as a claim about reality. Even if we have good general grounds for taking a particular data generating arrangement to be strictly deterministic, we in fact never know all the factors operating in any particular measurement. The combined action of these unknown factors produces a pattern of outcomes that is indistinguishable from that produced by a genuinely indeterministic system. Thus, from a practical point of view, we can simply treat the setup as being indeterministic. This form of "methodological indeterminism" at the experimental level is fully compatible with methodological determinism regarding theories, i.e., with the admonition always to seek a set of state variables which make the system deterministic. In fact, one can even espouse methodological determinism regarding experimental setups since we never do eliminate all unknown influences.

My view, then, is that theories are judged by the outcomes of statistical inferences regarding particular experimental systems. This apparently conflicts with much that has been written in recent years about the dependence of observation on theory.[11] Indeed, I seem to be advocating a version of the traditional three-level view of scientific knowledge and inquiry. Singular observations form the inductive basis for statistical hypotheses about experimental setups and these hypotheses in turn form the inductive basis for theories concerning broad classes of physical systems. In fact I will deny an important tacit assumption of this view, namely, that any statistical hypothesis can be inferred from singular statements *alone*. But this is not enough to avoid conflict with the more extreme advocates of 'theory-ladenness'. However, as I do not wish to debate this matter here, I will simply state my conviction that such views are wrong, and, moreover, that they are so because of a mistaken theory of meaning.

[11] Here, of course, I am referring to the vast literature which surrounds and to a large extent grew out of the work of Hanson (1958), Feyerabend (1962), and Kuhn (1962). I have dealt in a general way with some of this literature in Giere (1973b).

While rejecting views implying the impossibility of observationally testing scientific theories, I do agree with much that has been written about the conceptual and methodological importance of theory in scientific inquiry. In particular, the theories we hold may determine the *categories* of things we perceive without determining our *judgments* about them. In this sense theory is truly the midwife of observation. Moreover, it is generally the case that what a scientist investigates, and when and how, is a function of his theoretical commitments, vague though they may be. I give no hostages to that Popperian bogeyman, Baconian inductivism. Nevertheless, when it comes to validating those commitments before the scientific community, the logic to be employed follows the general pattern outlined above. That, at least, is the assumption that guides the rest of this study.

3. The Logic of Statistical Hypothesis Testing

As indicated above, there is no single, unified characterization of statistical inference that would be acknowledged in either the theory or practice of contemporary statisticians and scientists. In fact, the trend in recent years has been toward increasing eclecticism — in spite of fundamental incompatibilities among all the methods employed. In order to follow my general program, therefore, it will be necessary to depart from current statistical practice by isolating and reconstructing a single approach to statistical inference. The logic adopted here is a version of the theory of statistical hypothesis testing developed by Neyman and Pearson in a fundamental series of papers published between 1928 and 1936. This choice is based partly on my personal conviction that the Neyman-Pearson (N-P) viewpoint is basically correct. But it is also justified by the tremendous influence this work has had for nearly a half century. The N-P approach must count as one of the three or four major candidates for a comprehensive inductive logic. To produce a relatively complete and philosophically adequate reconstruction, however, would be a major undertaking. The following sketch is geared to the problem at hand, which is justification, not characterization, and is as simple as the demands of clarity and pedagogy will allow.[12]

[12] The basic original papers are now reprinted in Neyman and Pearson (1967). For a personal account of the formative years see Pearson (1966). Currently the definitive presentation of the theory is due to Lehman (1959). For an elementary exposition see Hodges and Lehman (1964). Hays (1963) has given a highly in-

The basic conception in my account is that of a *chance setup*.

CSU is a chance setup iff CSU is a physical process with one or more physically possible final states (or "outcomes").

This definition includes deterministic processes as a limiting case, but henceforth I shall be concerned only with systems having more than one physically possible outcome.[18] Let A_i be a possible outcome of CSU and U the set of all possible outcomes, i.e., $U = \{A_i\}$. Now let E_i be a subset of U and $F = \{E_i\}$ the family of all such subsets. It is these subsets of U, not the basic outcomes themselves, that mathematicians call 'events'. Following standard mathematical practice we shall assume there is a quantity $P(E_i)$, the probability of E_i, which obeys the standard axioms for probability spaces. For finite spaces:

(1) $P(E_i) \geqq 0$

(2) $P(U) = 1$

(3) $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ if $E_1 \cap E_2 = \Lambda$.

In keeping with my ontological preferences, I shall interpret $P(E_i)$ as the measure of the propensity of the physical system, CSU, to produce one of the outcomes constituting event E_i. And I assume that propensities operate on each individual trial of a CSU. Those who find such a conception excessively metaphysical or obscure may attempt to interpret $P(E_i)$ in terms of the relative frequency of event E_i in a (perhaps infinite) series of trials of the relevant CSU.

A precise characterization of statistical hypothesis testing requires a definition of a statistical hypothesis.

H is a statistical hypothesis iff H is a statement ascribing a distribution of probabilities to the outcomes of a CSU.

Taking a traditional example, conceiving a child might be considered a chance setup with the set of outcomes (male, female). The statement that $P(\text{male}) = .51$ is then a statistical hypothesis. Following standard mathematical practice we shall assume the existence of a real valued function (conveniently, but misleadingly, called a *random variable*) over the outcomes of any chance setup. A statistical hypothesis is then any statement describing the probability distribution over a random variable.

formed and sometimes critical presentation geared to applications in psychological research. My own exposition tends to follow Neyman (1950, now out of print).

[18] The term 'chance setup' has been borrowed from Hacking (1965). The usual term among statisticians is 'random experiment', but as the indicated processes may be neither random nor experiments, another terminology seems preferable.

For example, if X (male) $= 0$ and X (female) $= 1$, then $P(X = 0) = .51$ expresses the statistical hypothesis given above.

Many discussions of statistical inference take the concept of a *population* as fundamental. On such a view, a statistical hypothesis gives the relative frequency of a characteristic in some (perhaps infinite) population. The viewpoint above is more general and, I think, more fruitful. Populations may be dealt with in terms of a chance setup consisting of a sampling mechanism operating on the indicated population. How the distribution of population characteristics and the probabilities for different outcomes are related depends on the sampling mechanism, e.g., whether it is random or not.

On the Neyman-Pearson account it is crucial to distinguish simple from composite statistical hypotheses.

> H is a simple statistical hypothesis iff H specifies a unique probability distribution over a random variable.
>
> H is a composite statistical hypothesis iff H is a statistical hypothesis that is not simple.

The importance of this distinction will soon be apparent. We are now equipped to begin talking about *testing* statistical hypotheses. The intuitive notion of a statistical test is that one uses the results of a number of trials of the chance setup as evidence for or against the proposed statistical hypothesis. The N-P approach provides a precise rendering of this intuitive idea.

Imagine a series of n trials of a specified CSU, and let $P(X_j = x_j)$ be the distribution of probabilities over the jth trial. Representing each of the j random variables on a separate axis, the n trials determine an n dimensional *sample space* consisting of points representing each possible outcome of all n trials. That is, each point in the sample space is an n-tuple $x = (x_1, x_2, \ldots, x_n)$. Now let H be a simple statistical hypothesis ascribing a unique probability distribution to the outcomes of each trial. For example, H might give the *same* distribution to each trial. In this case there are well-known mathematical techniques for computing a probability distribution relative to H, $P_H(X = x)$, over the whole sample space. Continuing the standard example, the basic sample space for a series of n conceptions consists of the 2^n possible sequences of males and females. Given reasonable assumptions, however, it is sufficient for most purposes to consider only the sample space consisting of the $n + 1$ possible numbers of male offspring in n trials, i.e., $S =$

$\{0, 1, 2, \ldots, n\}$.[14] If $P(\text{male}) = p$ is the same on each trial, the probabilities of each of these possible sample points are given by the familiar binomial formula:

$$P(X = m) = \binom{n}{m} p^m (1 - p)^{n-m}.$$

Since at least the beginning of the eighteenth century, many authors have implicitly held that a statistical hypothesis is to be 'rejected' whenever the observed sample point would be highly improbable if that hypothesis were true.[15] But this intuition provides only a very incomplete characterization of a statistical test. In many cases *every* possible outcome is quite unlikely. For example, exactly 510 males in 1000 conceptions is highly unlikely even if $P(\text{male}) = .51$. Yet this result should provide greater support for the hypothesis than any other possible outcome.

By the second decade of this century, R. A. Fisher and others had focused on finding a *region* of the sample space (the 'critical' or 'rejection' region) for which the probability of occurrence is low if the test hypothesis is true.[16] If one resolves to reject the (simple) hypothesis under investigation if and only if the observed sample point falls in the rejection region, then there is only a small and controllable probability that we will by chance reject H when it is in fact true. This probability is called the *significance level* of the test of H, and it may be controlled by suitably adjusting the number of trials and the rejection region.

Though it is still a point of dispute, Fisher's characterization of hypothesis testing seems to me quite incomplete.[17] The most obvious gap is the lack of any systematic grounds for choosing among several possible rejection regions with the same significance level. Indeed, Fisher

[14] This reduction of the sample space is treated in discussions of 'sufficient statistics' found in any good text, e.g., Kendall and Stuart (1958–66).

[15] As early as 1710 Arbuthnot used data from the London registry to argue in effect that the hypothesis $P(\text{male}) = 1/2$ is to be rejected. He went on to argue, perhaps in jest, that this implies that sex determination is not a matter of chance and is therefore the result of divine intervention. For further details see Hacking (1965). Laplace also used an implicit theory of statistical hypothesis testing in his scientific work. Some examples are discussed in his popular *Philosophical Essay on Probabilities* (1814).

[16] The development of Fisher's views may be traced through his papers (1950). One should also read at least the early chapters of *The Design of Experiments* (1935). It is most illuminating to compare Fisher's discussion of "the lady tasting tea" (ch. 2) with Neyman's commentary (1950, ch. 5).

[17] Anscombe (1963), for example, argues that the basic Fisherian test of significance has important applications.

gives no systematic reason for not taking a small set of values near 510 males in 1000 trials as the rejection region for the hypothesis $P(\text{male}) = .51$. Moreover, though Fisher was careful to minimize the probability of accidentally rejecting a true hypothesis, he did not deal systematically with the problem of accepting a false one. This is in keeping with his statements that the only purpose of experimentation is to give the data a chance to reject a hypothesis.[18] But non-Popperians are not likely to be satisfied with only methods for rejecting hypotheses. An inductive logic that provides for rejection ought to provide for acceptance as well.

The developments initiated by Neyman and Pearson may be viewed as an attempt to fill in the gaps in Fisher's characterization of hypothesis testing. On their view, a test of H is to be characterized as a method of reaching a decision either to reject or to accept H. A good test is then characterized as one for which the probabilities of a wrong decision are not only low, but, more importantly, controllable. Now if both acceptance and rejection are possible, then there are not one but two kinds of possible 'mistakes'. There is the Fisherian case, or Type I error, of by chance obtaining a sample point in the rejection region even though H is true, and thus rejecting a true hypothesis. Conversely, one might obtain a sample point in the acceptance region even though H is false, and thus suffer a Type II error, accepting a false hypothesis.

If H is simple, the probability of Type I error is just the Fisherian significance level. But what about the probability of Type II error? Formally this is the probability of obtaining a sample point in the acceptance region if H is false. But since not-H is in general a *composite* hypothesis, this latter probability cannot be defined without a prior weighting of the component simple hypotheses. Let us postpone the question of how, or even whether, this is to be done, and simply assume that there are only two possible, logically exclusive, simple hypotheses, H and K. The probability of Type II error, i.e., accepting H if H is false, is then the probability, given K, of obtaining a sample point not in the rejection region of the sample space. The complementary probability, i.e., the probability of not making a Type II error is called the *power*

[18] "Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (Fisher, 1935, 4th ed., p. 16). I cannot help but speculate that Popper's sympathetic reception by British scientists may be partly the result of the widespread influence of Fisher's views. Gilles (1973) has recently attempted to develop a rule for rejecting statistical hypotheses along lines suggested by Popper (1959).

of the test. This is the probability that the test will detect the falsity of *H* if *K* is true.

We are now equipped to give a formal definition of a Neyman-Pearson statistical test.

> *T* is an N-P statistical test of hypothesis *H* relative to chance set-up CSU, sample space *S*, and alternative, *K* iff: *T* is an exclusive and exhaustive partition of *S* into two regions, one, *R*, designated as the rejection region.

A noteworthy philosophical consequence of this definition is that the N-P logic cannot be reconstructed as simply a logical relation between *H* and a description of the observed sample point. In addition to the CSU and sample space, there is essential reference to a set of *admissible hypotheses* which includes *H* and at least one simple, exclusive alternative hypothesis. The reference to alternative hypotheses, however, has more than general philosophical interest. It provides the means for giving a precise mathematical formulation of the problem of choosing an optimal rejection region. In some cases at least, this problem has a precise and intuitively appealing solution.

Continuing the case of testing a simple hypothesis against a simple alternative, it is natural to wish simultaneously to minimize the probability of both kinds of error, i.e., to minimize the significance level while maximizing power. Brief reflection, however, shows that for a fixed number of trials, these two probabilities are inversely related. Thus the statement of the problem given above is not well formed. Suppose, then, that we fix the significance level at some conveniently small value, say .05, and then seek to maximize the power of the test for a fixed number, *n*, of the trials. In this case one obtains the following elegant result: Let R consist of all those sample points, x_i, of S for which

$$P_H(X = x_i)/P_K(X = x_i) \leqq c,$$

where $c > 0$. The significance level of this test is then

$$P_H(X \epsilon R) = \alpha,$$

where α can be chosen arbitrarily by a suitable choice of c. It is then fairly easy to show that there is no other rejection region, *R′*, for which $\alpha' \leqq \alpha$ and for which the power of the test is greater. In sum, the likelihood ratio of points in the sample space may be used to determine a rejection region which has the minimum probability for Type II error relative to a fixed maximum probability for Type I error. More briefly, the likelihood ratio determines a test which has maximum power for a fixed maximum significance level.[19] This test clearly accords well with the basic intuition that one can confidently reject a hypothesis in favor of an alternative if the observed result is much more probable on the alternative (even though both probabilities may be small).

Now suppose that the alternative hypothesis is composite, e.g., let $K = K_1$ or K_2, where K_1 and K_2 are both simple statistical hypotheses. As long as the test hypothesis H is simple, the significance level relative to any rejection region is well defined, i.e., $\alpha = P_H(X \epsilon R)$. Being a weighted sum of two probabilities, however, the overall power $P_K(X \epsilon R)$ is not determinable without further assumptions. That is:

$$P_K(X \epsilon R) = w_1 P_1(X \epsilon R) + w_2 P_2 (X \epsilon R).$$

The obvious interpretation of w_1 and w_2 is as the prior probabilities $P(K_1)$ and $P(K_2)$. This interpretation, however, is difficult to square with an objective, empirical interpretation of probability, except in those rare cases when a super CSU produces ordinary CSU's as outcomes. And even in these cases the probability distribution of the super CSU may not be known. For these reasons it is desirable to find a method which requires no prior probabilities for any hypotheses.

In recent years, of course, prior probabilities, in the form of coherent subjective degrees of belief, have become much more respectable than they once were. But if one adopts this view, the natural course is to characterize inductive inference as a subjective probability relation and to apply Bayes's theorem. There is no need for an independent account of hypothesis testing. Thus any appeal to subjective prior probabilities of hypotheses in reconstructing a Neyman-Pearson account of hypothesis testing would take us out of the Neyman-Pearson tradition altogether. Of course the Bayesian approach might in the end prove the more adequate, but that is another question.

The Neyman-Pearson response to composite alternatives is to avoid any talk about the overall power of a test. Instead, such a test is char-

[19] The above is an elementary statement of "The Fundamental Lemma of Neyman and Pearson," first proved in their 1933 paper "On the Problem of the Most Efficient Tests of Statistical Hypotheses." Reprinted in Neyman and Pearson (1967). Lehman (1959, ch. 3) gives a detailed proof and Hodges and Lehman (1964, pp. 351–52) an elementary version. My statement of the lemma is strictly correct only for continuous distributions, but this need not bother us here.

acterized by the power viewed as a function of a parameter which characterizes the set of simple alternatives. In certain cases this function has obviously desirable properties. For some sets of admissible hypotheses, for example, it is possible to find a rejection region which is simultaneously most powerful for each simple alternative in *K*, given a fixed significance level. Thus even though the power of the test varies for different simple components of the composite alternative, for no simple alternative is there a more powerful rejection region at the same significance level. Such regions are called Uniformly Most Powerful (UMP) regions. For example, the likelihood ratio criterion determines a UMP region for a test of $P(\text{male}) = .51$ *vs* $P(\text{male}) > .51$.

When a UMP region exists, it determines the best possible test for any number of trials. For many composite alternatives, however, there is no UMP region, i.e., $P(\text{male}) = .51$ *vs* $P(\text{male}) \neq .51$. In such cases Neyman and Pearson introduce a *new* condition on optimal rejection regions, namely, for all $K_i \epsilon K$,

$$P_i(X \epsilon R) \geqq P_H(X \epsilon R).$$

This condition, called "unbiasedness," simply means that one is always more likely to reject the test hypothesis if it is false than if it is true. Often, as in the example above, there is a UMP region among the unbiased regions which is then taken as optimal. But while this condition is plausible, though hardly self-evident, there are sets of admissible hypotheses for which not even UMP unbiased tests exist.

Over the years other criteria for determining various classes of UMP tests have been proposed.[20] We need not consider them here. Two points, however, are important. First, the notion of a uniformly most powerful test largely preserves the idea of building a method of accepting and rejecting hypotheses which has known and controllable error probabilities. Second, the theory of testing simple hypotheses against composite alternatives is still open-ended. Depending on the set of admissible alternatives, one could introduce new criteria for optimal rejection regions in order to obtain a UMP test of some kind. This move is always open.

Turning finally to cases with both a composite test hypothesis and a composite alternative, it is clear that neither the significance level nor the power is uniquely determined without a distribution of prior prob-

[20] See, for example, Lehman's chapter on "invariant" tests (1959, ch. 6).

abilities over the component simple hypotheses. For such cases Neyman suggests an obvious extension of the methods used for simple hypotheses. Consider the ratio

$$\lambda(x) = P_i(X = x)_{max}/P_j(X = x)_{max}$$

where x is a particular sample point, H_i and K_j are component simple hypotheses of H and K respectively, and $P_i(X = x)_{max}$ is the least upper bound of the probability $P_i(X = x)$ for components H_i of H. The rejection region is then to consist of all sample points for which $\lambda \leq c$, where c is some small fraction.[21] The basic idea, of course, is that H is to be rejected if there is any simple alternative for which the observed outcome is much more likely than it is on any component of H. But here the rationale in terms of known error probabilities applies even less well than in the previous case. There at least the probability of Type I error was well defined. Here one is in effect reduced to comparing the *most favorable* member of *each* of the two sets of hypotheses. Lacking prior probabilities this is a reasonable compromise strategy, but it is all based on possible rather than actual error probabilities. For these reasons it is not surprising that one finds little literature on the development or application of such tests.

This completes my sketch of the logic of statistical hypothesis testing. Before turning explicitly to questions of epistemological justification, however, there are several points about the logic that require mention. The first is that, granting only some general mathematical conditions, any contingent statistical hypothesis can be tested by some N-P test. This follows because any statistical hypothesis must be either simple or composite, and one can always find some alternative, simple or composite, to form a test. One could even effectively avoid the λ-criterion by always choosing a simple alternative. Let us summarize this result by saying that the N-P logic is *weakly comprehensive*. The logic is also *strongly comprehensive* if the λ-criterion is regarded as fully legitimate. That is, for any H and any K, both either simple or composite, there exists an N-P test of H against K. But strong comprehensiveness does not hold if one is restricted to a fixed narrower class of tests, e.g., UMP

[21] Note that if H and K are both simple, the λ-method reduces to the likelihood ratio test described above. My statement of the λ-criterion differs from Neyman's in that his denominator refers to the whole set of admissible hypotheses, not just the alternative hypothesis. This difference is inconsequential here. See Neyman (1950, pp. 340–42).

unbiased tests. The general comprehensiveness of the N-P logic of testing will play an important role in the epistemological justification to be developed in section 7.

Finally, it must be acknowledged that the characterization of statistical hypothesis testing described above is itself incomplete in at least one important respect. For Neyman and Pearson, testing a hypothesis is a systematic way of reaching a decision to accept or reject that hypothesis. But what is it to decide to accept (or reject) a hypothesis? In his later writings, though not necessarily in his statistical practice, Neyman identifies accepting a hypothesis with choosing a course of action.[22] A statistical test thus becomes, in Wald's terminology, a statistical decision rule. For each possible outcome of a set of trials, the rule prescribes a particular course of action. Viewing hypothesis testing as a special case of decision theory not only tells us what "accepting a hypothesis" means, it also gives us a way of determining specific values for the significance level, power and number of trials for a test. The logic as characterized above only shows how these quantities are related; it does not determine an optimal set. In a decision theory context, these are in effect determined by the gains and losses associated with the possible results of the decision, i.e., a specific action in a certain actual situation.

Despite considerable initial plausibility, the Neyman-Wald approach has severe liabilities. It leads to the unresolved and perhaps unresolvable problem of finding a general criterion for decision-making under uncertainty. Minimax is not an adequate general strategy. Moreover the kinds of utilities typically found in practical decision-making contexts are nonexistent, or at least exceedingly hard to find, in scientific research contexts.

An alternative to a full-fledged decision theoretic approach is to take "accepting *H*" to mean something like adding *H*, tentatively, to the body of scientific knowledge. The choice of significance level and power would then turn not on anything like monetary profits and losses, but on the desire to increase the body of knowledge in content, precision, etc., as efficiently as possible. Hopefully one could use this goal to justify the decision rule implicit in at least some N-P tests without appealing to

[22] I take this to be a consequence of Neyman's concept of 'inductive behavior'. See his " 'Inductive Behavior' as a Basic Concept of Philosophy of Science" (1957). Pearson, by the way, is reputed to take a much less behavioristic view of acceptance.

any completely general strategy like minimax. Some work has been done along these lines, but much more is necessary before the logic of testing presented above could be suitable as part of an adequate overall account of inductive reasoning.[23]

4. The Epistemological Structure of Statistical Tests

Suppose that a statistical hypothesis, *H*, has been accepted (or rejected) as the result of an N-P test, *T*.[24] In this case the acceptance of *H* is *validated* by *T*. There are, however, several legitimate ways one could question whether the acceptance of *H* is "really" justified in some more ultimate sense. One could, for example, question whether the acceptance of any hypothesis should depend solely on the supposedly favorable error probabilities of an N-P test. But this question concerns the adequacy of the whole N-P approach and will not be considered here. Or, one might question whether the significance level and power were sufficiently stringent. But this question again concerns aspects of the N-P characterization which cannot be considered here. However, even granting that *T* represents an appropriate application of N-P methods, one might still question the acceptance of *H* on the grounds that the stated error probabilities may not in fact be correct. On what grounds, then, may it be maintained that the indicated error probabilities are indeed correct? This question gives rise to a regress typical of Humean arguments against the justifiability of induction.

Looking back at the characterization of N-P tests it is fairly easy to see that in general there are two, and only two, points at which *empirical* assumptions may enter into the formulation of an N-P test. One is due to the necessity for designating a set of admissible hypotheses since it is assumed that some member of this set, i.e., either *H* or *K*, is true. Thus only if the disjunction, *H* or *K*, is *logically* exhaustive will the need for a set of admissible hypotheses fail to yield an empirical assumption. A second possible source of empirical assumptions opens up whenever a test involves more than one trial of a CSU. Any claim of equality

[23] In a forthcoming paper, Giere (1975), I develop this approach both for tests of statistical hypotheses and for tests of theories. This later paper thus provides an answer to the arguments of Carnap and Jeffrey (cited in fn. 3 above) that it is impossible to incorporate 'acceptance rules' into an adequate inductive logic.

[24] In cases where no confusion should arise I will simply say "accepted" rather than saying "accepted or rejected."

(or difference) of probability distributions on two or more trials is an empirical claim. All but a very specialized class of N-P tests incorporate such claims.

I will refer to assumptions of the first type as *admissibility assumptions*. Assumptions of the second type will be called *multiple trial assumptions*. Finally, I will refer to any empirical assumption of an N-P test as a *presupposition* of that test. This term is natural and appropriate since the truth of the presuppositions is assumed whether the test hypothesis is accepted or rejected.[25]

We are now in a position to recognize a feature of the N-P characterization that has important consequences for the justification problem, namely, that all presuppositions of an N-P test are themselves statistical hypotheses. That admissibility assumptions are statistical hypotheses is obvious. Indeed, it is clear that the admissibility assumption of any N-P test is just the *composite* statistical hypothesis, *H* or *K*. That multiple trial assumptions likewise include only statistical hypotheses is less obvious. Given multiple trials, there are just two questions that must be answered if they are to be part of a single test. The first is whether the distributions of probabilities are the same or different on each trial. The second is whether the trials are independent. A statement of sameness or difference of distributions on a series of trials is just a conjunction of statistical hypotheses regarding the individual trials and thus itself a statistical hypothesis. To see that independence of trials also involves only statistical hypotheses, let X and Y be the random variables associated with two trials of a CSU. The case of more than two trials is more complex, but not in principle any different.[26] Now two trials are said to be independent iff

$$P(\mathrm{X} = \mathrm{x}\ \&\ \mathrm{Y} = \mathrm{y}) = P(\mathrm{X} = \mathrm{x})\ P(\mathrm{Y} = \mathrm{y})$$

for all x and y. Stated in terms of conditional probabilities this condition becomes

$$P(\mathrm{X} = \mathrm{x} \mid \mathrm{Y} = \mathrm{y}) = P(\mathrm{X} = \mathrm{x}).$$

[25] From now on it will be assumed that *all* empirical presuppositions of an N-P test are either admissibility or multiple trial assumptions. I am thus ignoring the fact that performing an N-P test requires such things as the observation and recording of data which are possible only under certain empirically given conditions. This seems perfectly legitimate in a study of the finer structure of scientific inference.

[26] For the definition of independence for more than two trials see any good textbook of probability or statistics.

It is clear that these conditions are just complex statistical hypotheses relative to the joint CSU consisting of two trials of the original CSU. Standard tests for such hypotheses are discussed in many texts and related tests will be treated in the following section. The intuitive idea, of course, is that the probability distribution for any one trial should have no influence on the distribution of any other.

Recall, now, our conclusion that the N-P characterization is comprehensive in the sense that any statistical hypothesis can be subjected to some N-P test. Given that any presupposition of an N-P test is a statistical hypothesis, it follows that any presupposition of an N-P test can itself be tested by an N-P test. This result may be summarized by saying that the N-P characterization of statistical hypothesis testing is *presuppositionally closed*.[27]

Let us now return to the situation in which the justifiability of accepting H as a result of performing test T is questioned on the grounds that some presupposition, P, of T may not be true. Given presuppositional closure, it is always possible that this question be answered without appealing to any other inductive method for testing statistical hypotheses. P may itself have been N-P tested and accepted. But here the possibility of a Humean regress sets in. In general the test T' of P will itself have presuppositions which may in turn be questioned, and so on. Thus, assuming we have only the N-P logic for testing hypotheses, at any particular time the ultimate basis for the acceptance of H has the sort of rootlike structure shown in the diagram on page 234, Figure 1. I will call P_1 and P_2 the *immediate* presuppositions of T. All others are *indirect* or *remote* presuppositions of T. (Note that P_3 and P_4 are immediate presuppositions of T'.) In the case of presuppositions like P_1 and P_4 which have themselves been tested, it is assumed that the tests were positive, i.e., P_1 and P_4 were accepted. This point will be discussed further in section 7. Here I will only note that this picture of the ultimate justifiability of accepting H does not presume the erroneous principle of the transitivity of conditional probabilities. We are not here arguing that P_4 makes H probable because P_4 makes P_2 probable and P_2 makes H probable. In the N-P logic there is no relation of conditional probability between the presuppositions of a test and the test

[27] I owe this apt term to Lewis Creary who commented on an ancestor of this paper at a session of the May 1970 meeting of the American Philosophical Association, Western Division, in St. Louis.

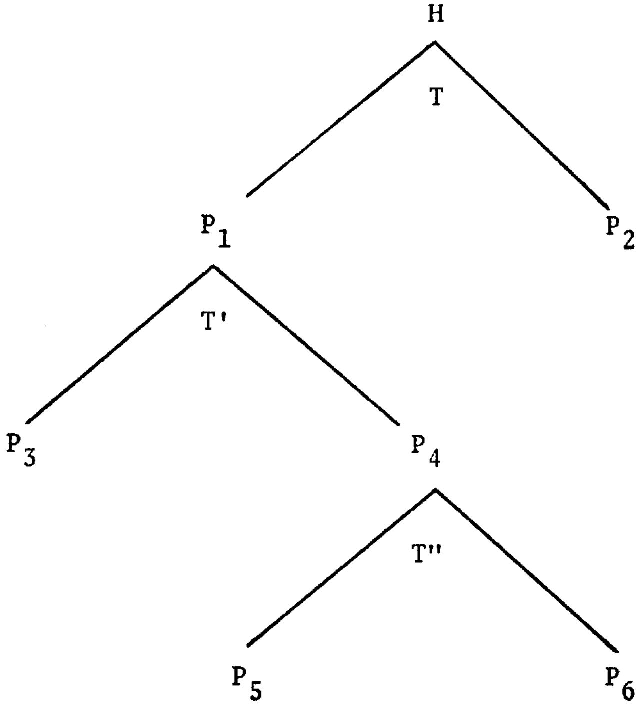

Figure 1

hypothesis. On the other hand, I could hardly claim that there are no further relations among hypotheses like P_1, P_4, and H that need further investigation.[28]

The root structure of statistical tests illustrated above suggests two different general programs for constructing an ultimate justification for accepting a statistical hypothesis. Both focus primarily on those presuppositions currently not connected to further presuppositions. Let us call these the *ultimate* presuppositions of T, notwithstanding the fact that the set of ultimate presuppositions of any N-P test may change from time to time. The first program, which follows what I call the 'foundationist' tradition, is to "seal off" the root structure by insisting that all ultimate presuppositions be subjected to N-P tests whose presuppositions are purely logical. The alternative program is to find an ultimate

[28] Indeed, all the questions about the applicability (practical or otherwise) of 'accepted' hypotheses may be raised here. There is also the possibility of a lottery paradox. For references see Kyburg (1970) and fns. 3 and 23 above.

justification allowing ultimate presuppositions which are so simply because they have not yet been tested and therefore have been neither accepted nor rejected. I will develop a "methodological" version of this approach in section 7.

5. A Semirealistic Example

An example may be helpful in conveying some intuitive feel both for the logic of testing and for related epistemological issues. One must be wary, however, of simple textbook examples which, by excessive abstraction from any realistic context, often distort more than they illuminate. The ideal example would be a miniature case study of a scientific inquiry. The following "semirealistic example" falls between the excessively abstract and the ideal.

Since ecology became a subject of public concern there has been increased sensitivity to questions about the effect of various substances on human and other organisms. A fairly crude example of such a question would be: Do certain low levels of DDT decrease the adult size of mammals? Of course a scientific study of this question would require precise criteria for size, DDT level, and other variables, but let us ignore such details.

I would regard the claim that DDT decreases size as a claim about a kind of probabilistic physical system, e.g., that exemplified by a particular strain of laboratory rat. At birth each animal has a propensity distribution over possible adult sizes. The general claim under question may be that the mean of this distribution is a decreasing function of DDT levels. This is not merely a statistical hypothesis but a theoretical function relating state variables of a probabilistic physical system to a probability distribution over other states. But the theoretical hypothesis has statistical consequences, for example, that the expected, i.e., most probable, adult size of rats exposed to DDT will be lower than that of rats otherwise similar but not so exposed. This kind of hypothesis is subject to direct statistical tests.

Before looking in more detail at the appropriate statistical tests, one should note the methodologically necessary but epistemologically minimal role played by high-level theory in cases like this. At present, biological theory is not so well developed as to yield the unequivocal prediction that DDT will tend to lower the adult size of rats. Still less can any theory yield a precise form for the functional relation between

DDT exposure and adult size. Yet existing theory is important because it tells us that DDT is the kind of substance that *could* lower the expected adult size of mammals. Without this minimal theoretical background it is very unlikely that the question of such a relationship would arise or be investigated in a serious scientific context. And with good reason. Scientific inquiry cannot efficiently be guided by consideration of what is merely logically possible. It is usually necessary to limit the field to what one regards as physically possible.

Turning to the problem of designing a statistical test, imagine two groups of rats of the appropriate type. The members of one, the experimental group, are subjected from birth to specified levels of DDT; the other group, the controls, are kept free of DDT. Now the hypothesis that the expected size of the experimental rats is lower than that of the controls is a composite hypothesis and thus not an ideal subject for statistical testing. On the other hand, the hypothesis that there is no difference in expected means is simple and thus better suited to our logic. To be more precise, let the expected sizes of the experimental and controls be μ_E and μ_C respectively. The test hypothesis, or 'null hypothesis', then, is H: $\mu_C - \mu_E = 0$. Realistically there are only two alternative hypotheses that might be considered, i.e., $\mu_C - \mu_E \neq 0$, and $\mu_C - \mu_E >$ 0. Which we use depends on whether or not our previous knowledge is sufficient to reject the possibility that DDT increases the expected adult size of rats. Assuming such knowledge we have K: $\mu_C - \mu_E > 0$.

In addition to a set of admissible hypotheses we require a suitable sample space and rejection region. Let us assume that at birth the probability distribution of each rat over possible adult sizes is normal with mean μ_C or μ_E and variance σ. In short, the only possible difference in distributions is that the experimentals may have their curves shifted downward. Finally, we assume that all distributions are unaffected by the distribution for any other member of the group. With these assumptions the standard sample space would be all possible values of the difference $d = m_C - m_E$ of the *observed* mean adult sizes of rats in the two groups. Now suppose there is some difference $\delta_0 > 0$, which, for theoretical or practical reasons, is considered important. By requiring a fairly high power for this particular alternative, say .8, we determine the required sample size n and rejection region $d > d_0$. The resulting test, the well-known 't-test', is UMP for H against K.[29]

[29] The t-test is discussed in any text of applied statistics, e.g., Hays (1963).

The presuppositions of this test are apparent. The admissibility assumptions are (i) that $P(X)$ is normal with mean μ_C or μ_E and fixed variance σ, which need not be known, and (ii) that $\mu_C - \mu_E \geqq 0$. The multiple trial assumptions are (i) that $P_E(X)$ is the same for all experimentals and that $P_C(X)$ is the same for all controls, and (ii) that all distributions are independent of all outcomes.

Anticipating upcoming epistemological questions, let us consider briefly how one might justify some of these presuppositions. In practice one would obtain experimental animals from a laboratory that specializes in breeding genetically homogeneous strains of rats. That each rat has very nearly the same expected adult size and the same reactions to various chemical substances would then be justified both by genetic theory and by long experience in breeding that particular strain. Epistemologically, however, one might object to the extent to which this justification relies on a high-level theory which itself must have been validated through previous statistical tests.

There is a common procedure which both decreases reliance on high-level theory and also eliminates any direct ontological commitment to propensities of individual organisms. The procedure is to isolate a large population of similar rats. This may require some theoretical knowledge, but not nearly so much as breeding because it is not required that each rat have the same propensities. Rather, all that matters is that each rat have a definite adult size and that the distribution of these sizes in the population has a known form, e.g., normal. Whether the individual rats are deterministic systems or not is immaterial. The CSU necessary to determine a statistical test is introduced in the form of a mechanism which selects samples from the indicated population. It is convenient to think of the sampling mechanism as having propensities for selecting different members of the population. The simplest assumption is that the mechanism is perfectly random, i.e., each member of the population has the same propensity of being selected on each draw. A weaker but still sufficient assumption is that even if the sampling mechanism favors some rats over others, its biases are independent of the attained adult size of any rat. In either case, each trial of the sampling mechanism will have a propensity distribution over adult sizes that exactly matches the distribution of sizes in the population. Thus we need only select two samples of n rats, subjecting one to the specified DDT level, and the test proceeds as above. As usual, those who prefer not to think in terms of propensities may attempt another analysis of random sampling.

Suppose the above test results in a decision to reject the null hypothesis. Strictly speaking, all that has been decided is that the observed difference in average adult size indicates an important real difference in expected size and is not just an extreme statistical fluctuation. Assuming that the decision is in fact correct, it does not logically follow that the reduction in expected size was *caused* by DDT. This would follow only if *all* other relevant variables were the same for both groups. But in practice we do not even know what all the relevant variables are, let alone that they have all been held constant. It might be, for example, that the experimental animals were handled more in the laboratory than the controls, and that handling alone retards growth by the observed amount. Of course we try hard to keep every possible relevant factor constant, and here random sampling and random assignment help. But such procedures do not guarantee success.

Here we see, therefore, an example of the inferential gap that in practice always exists between statistical conclusions and theoretical or causal hypotheses. Unfortunately I must once again decline to discuss the detailed logic of the second-level inference. I only wish to emphasize that there are two levels and that failure at the second level does not necessarily indicate any defects in the characterization or the application of the lower level statistical logic.

6. The Rejection of Foundationist Justifications for Hypothesis Testing

Much of the twentieth century literature on the justification of induction begins with the tacit assumption that the goal is to find a justifiable method of ampliative inference which requires only singular observation statements as premises. Only such a method, it is held, could avoid the ultimate epistemological circularity of an inductive logic which required as premises statements that could themselves only be justified by an ampliative inference. Following Nicod (1930) and Reichenbach (1949), two authors in this tradition who were very much aware of their goals, I will refer to the desired form of inference as a *primary* inductive method. All others are *secondary*.

With a primary inductive method at hand one might hope to build all knowledge on a foundation of singular facts only. This is certainly what Reichenbach tried to do with his rule of simple enumeration. Let us now see whether this 'foundationist' program could be carried out within the framework of N-P hypothesis testing. It is clear that to complete such a program there must be some bona fide N-P tests which make no nontrivial admissibility or multiple trial assumptions. Let us therefore investigate the possibility of constructing such a test.[30]

The simplest distribution is the binomial. If there is no presupposition-free N-P test for a binomial parameter, then there are no such tests at all. Continuing the example of the previous section, pick a size S_o for adult DDT-free rats. Then the adult size of each such rat has two possible values, $x \leqq S_o$ or $x > S_o$, and $P(X \leqq S_o) = 1 - P(X > S_o)$. Assuming n trials, consider the simple statistical hypothesis:

H: (i) $P(X_i \leqq S_o) = 1/2$ for all trials.
(ii) All trials mutually independent.

For ease of exposition I have made the needed multiple trial assumptions part of H itself so that all presuppositions of the test may be treated as admissibility assumptions. If we are to avoid all contingent presuppositions, the alternative for our test of *H* must be the composite hypothesis:

K: not-*H*.

The union, *H* or *K*, is therefore logically exhaustive and the test has only logically true presuppositions. The question is whether with such a broad class of admissible hypotheses one can construct a bona fide N-P test.

The basic sample space underlying any test of *H* vs. K would be the set of 2^n n-tuples $x = (x_1, x_2, \ldots, x_n)$ where x_i has two possible values, say 0 or 1. Now let *R* be any possible rejection region, i.e., *R* is any nonempty and nonexhaustive subset of n-tuples in the basic sample space. Given the inclusiveness of *K*, it is easy to prove that there is some $K_j \epsilon K$ for which $P_j(X \epsilon R) = 0$. Since R does not exhaust the sample

[30] Both Reichenbach (1949, pp. 360–61) and Pap (1962, p. 238) have explicitly claimed that N-P testing is a secondary inductive method and therefore incapable of being a philosophically fundamental inductive logic. Neither author, however, gives a detailed argument why this should be so. Such an argument has recently been given by Rogers (1971). I am grateful to Professor Rogers for prodding me into thinking more deeply about this question.

space, there is some sample point $x' \notin R$. Suppose this point is $x' = (x_1 = 1, x_2 = 0, \ldots, x_n = 0)$. Now since K contains all possible hypotheses specifying unequal probabilities on different trials, there is a $K_j \epsilon K$ for which $P_j(X_1 = 1) = 1$, $P_j(X_2 = 0) = 1, \ldots, P_j(X_n = 0) = 1$. Thus, $P_j(X \epsilon R) = 0$.

It follows immediately that R does not provide a UMP test of H vs. K since there obviously exists another possible rejection region with the same or lower significance level, namely $R' = x'$, for which the power against some alternative, namely, K_j, is greater. It also follows that R does not provide a UMP unbiased test of H vs. K since for finite samples the significance level $P_H(X \epsilon R)$ must be greater than zero and unbiasedness requires that the power be greater than the significance level for all simple alternatives. (The definition of unbiased tests was given in section 3.) For most statisticians this would be sufficient proof not only that there is no "best" N-P test of H vs. K, but also that there is not even a "good" N-P test. In practice one might tolerate a test that is slightly biased for a narrow class of alternatives. But in principle it is hard to countenance a test with many alternatives for which one is more likely to reject the null hypothesis when it is true than when it is false. This is the case with the test above.

It is, of course, not logically impossible that someone should propose an intuitively satisfactory criterion for good though clearly biased N-P tests like the above. As noted in section 3, the theory of testing simple against composite hypotheses is still somewhat open. In the absence of a definite proposal, however, one need not take this possibility very seriously. Moreover, I am confident that most students of statistical theory would judge it almost inconceivable that there should be an appealing criterion that could be satisfied in the face of so broad an alternative as that in the test above.

Let us grant, therefore, that there is little hope of constructing a legitimate empirically presuppositionless N-P test of a simple hypothesis versus a composite alternative. It might be suggested, however, that one can do better testing a composite hypothesis against a composite alternative. This approach has the disadvantage that the λ-criterion is the least well motivated of all N-P test criteria. Still, the possibility of using such tests in a foundationist program is worth investigating.

Let the test hypothesis assert "random trials," i.e.,

H: (i) $P(X_1 \leqq S_o) = P(X_2 \leqq S_o) = \ldots = P(X_n \leqq S_o) = p$.
(ii) All trials mutually independent.

The logically exhaustive alternative is K: not-H. H is composite because p may have any value between 0 and 1. In this case the rejection region is determined by the lambda principle, i.e.,

$$\lambda(x) = P_i(X = x)_{max}/P_j(X = x)_{max} \leqq c.$$

Recalling the definition of K, it is easy to see that $P_j(X = x)_{max} = 1$ for any x. No matter what the pattern of 0's and 1's in a sequence of trials, there is a $K_j \epsilon K$ which has $P_j(X = 0) = 1$ for all trials with outcome 0 and $P_j(X = 1) = 1$ for all trials with outcome 1. Thus the value of $\lambda(x)$ depends solely on the value of $P_i(X = x)_{max}$. Here we run into difficulties. Assuming n even, the lowest value of $P_i(X = x)_{max}$ occurs whenever there are $n/2$ 0's, regardless of order, and H_i is

H(½): (i) $p_1 = p_2 = \ldots = p_n = ½$.
(ii) All trials mutually independent.

Thus the minimal nonempty rejection region contains all outcomes with $n/2$ 0's, and no others. As far as the formalities of the lambda rule are concerned, we have here a presupposition-free test. But in terms of the general guiding principles of the N-P approach, this test is completely unsatisfactory.

Suppose $H(½)$ is true and a series of n trials results in the most probable outcome, i.e., half 0's and half 1's. The test described above would reject the hypothesis. Even if the 0's and 1's were ordered in an intuitively random pattern, any pattern, the hypothesis would be rejected. On the other hand, suppose $P_j(X = 0) = 1$ for the first $n/4$ trials and $P_j(X = 0) = 0$ for the ¾ n trials. Given the outcome that this alternative hypothesis is certain to produce, the hypothesis of equal, independent distributions would *not* be rejected. In short, even though we cannot appeal explicitly to significance level and power in tests involving two composite hypotheses, it is clear that this particular test is not very good at distinguishing random trials from nonrandom trials. And this will remain true no matter how we enlarge the rejection region or reduce the sample space.

I conclude that there simply are no satisfactory empirically presuppositionless N-P tests. The foundationist program for justifying the logic of statistical hypothesis testing must, therefore, be rejected. The underlying reason for the nonexistence of legitimate presupposition-free tests should be evident. A logically exhaustive set of admissible hypotheses is just too broad to permit meaningful discrimination on the basis of empirical trials alone. A legitimate application of the N-P logic of hypothesis testing requires a more narrowly prescribed set of admissible hypotheses, and this requires empirical presuppositions.

7. Methodological Justification: The Epistemological Roots of Statistical Knowledge

Empiricist epistemology has been dominated by the foundationist viewpoint since at least the seventeenth century. For those believing in the possibility of rational inductive reasoning, foundationism leads immediately to the search for a justifiable primary inductive method. Thus, abandoning the attempt to ground the acceptance of statistical hypotheses on presupposition-free tests means abandoning the foundationist program altogether. There is, however, another possible approach, one which emerges explicitly in the writings of Peirce, Dewey and, more recently, Karl Popper.[31] According to this 'methodological' viewpoint, no primary method is needed. It is enough that our methods of validation permit the process of scientific investigation to proceed in an "appropriate" fashion. Thus the ultimate justification of scientific hypotheses is not determined by a relation to any "foundation," but by rules governing the processes by which the whole body of accepted scientific claims changes and grows.

The task before us, then, is to mold this general conception of a methodological justification into a detailed account of the ultimate grounds for accepting the results of standard N-P tests. As explained in section 4 above, underlying an N-P test is a root structure of presuppositions, some of which have been subjected to tests introducing their own further presuppositions. At any time the root structure for

[31] Peirce's methodological writings are found throughout his *Collected Papers* (1934–58), but especially in vols. 2 and 5. Dewey's most systematic exposition is in his much-neglected *Logic: The Theory of Inquiry* (1939). The essays in *Conjectures and Refutations* (1962) form the most accessible introduction to Popper's philosophy. Given Popper's dogmatic insistence that there is no such thing as inductive reasoning, he could hardly be expected to support my use of some of his ideas (which, however, are mostly to be found in Peirce anyway).

any particular tested hypothesis is finite, though in principle it could be extended indefinitely. Furthermore, since there are no completely presupposition-free tests, the roots of a given test will necessarily terminate at untested "ultimate" presuppositions. We shall thus be constructing an ultimate justification for accepting a statistical hypothesis in terms of methodological rules largely concerned with the ultimate presuppositions of an N-P test. It must be emphasized, however, that the rules to be developed give conditions which the presuppositions must satisfy in order that the acceptance (or rejection) of *H*, the tested hypothesis, may be fully justified. These rules in no way constitute a primary inductive method leading to the acceptance of the presuppositions themselves. The ultimate presuppositions, in particular, may remain untested and therefore be neither accepted nor rejected.

In practice, some and possibly all presuppositions of a particular test may be indirectly well supported by theories already confirmed through previous statistical inquiries. But if one holds, as I do, that the ultimate epistemological grounds for accepting a theory are eventually all statistical conclusions, the appeal to theories can play no essential role in an inquiry into the ultimate justifiability of scientific claims. The following discussion, therefore, ignores the role of theories in the context of ultimate justification.

Before proposing specific methodological rules, we must consider a prior requirement on the rules themselves, namely, that the applicability of the rules be ascertainable without assuming the truth of any statistical hypotheses. The basis for this requirement is the simple desire to avoid the regress involved in using methodological rules to provide ultimate justification for accepting statistical hypotheses while simultaneously assuming the truth of statistical hypotheses in applying the rules. This requirement, which derives from Hume's analysis of induction and is assumed in all foundationist programs, can be defended on more positive grounds as follows. We can never be as certain of the truth of statistical (or universal) hypotheses as we are of singular statements which describe individual events. The most we can do is to find methods of utilizing singular statements to make it "reasonable" or "rational" to accept statistical hypotheses. But at least we should be certain that we are being rational in accepting *H*, even though we cannot be certain that *H* is true. But this is not possible if in applying the criteria for rational acceptance we assume the truth of statistical hypotheses. For example, we cannot say that the acceptance of *H* is (ultimately) justified if the immediate presuppositions of the relevant test are true. Since we can be no more certain of the truth of these presupposed statistical hypotheses than we can be of the truth of *H*, we could be no more certain that our acceptance of *H* was justified than that *H* is true. This much of the foundationist tradition seems to me correct and I will adopt the Humean requirement stated above.[32] Note, by the way, that we are not making the disastrous demand that to know *H* one must know that one knows *H*. We demand only that to be justified in accepting *H* one must know that he is so justified.

Turning now to consideration of the desired methodological rules themselves, a very plausible rule for justifying the acceptance of *H* relative to test *T* is:

> (1) No presupposition, *P*, of *T* shall have been rejected by an appropriate statistical test.

This rule is directed only at the tested presuppositions of *T*, if any, since ultimate presuppositions are by definition untested and thus neither accepted nor rejected. It simply reflects the fact that the reasonableness of accepting *H* through *T* rests on the calculated error probabilities of *T* and that these depend on the truth of *T*'s presuppositions. To accept the falsity of *P*, therefore, is to admit that the error probabilities of *T* have been calculated incorrectly and thus to destroy the grounds for accepting *H*. This does not mean, of course, that the calculated error probabilities might not accidentally be correct after all — but justification cannot depend upon happy accidents.

Note that rule (1) only says that if *T′* is a test of presupposition *P* in the root structure underlying the acceptance of *H*, then the result of *T′* must have been positive. It does not say, on pain of regress, that the acceptance of *P* must itself by ultimately justified. It will turn out, however, that if the acceptance of *H* is ultimately justified, then so is the acceptance of any tested presuppositions of *T*.

Let us focus now on the ultimate (i.e., untested) presuppositions of a statistical test. The remaining conditions constitute an attempt to

[32] Even if the Humean requirement were to be rejected, it would still be worth knowing how far one can get without crossing this line because then we might discover precisely why and how the line must be crossed. In this and other matters I may be accused of being too much influenced by my staunchly foundationist former colleague Wesley Salmon. Perhaps so, but these views are mine nevertheless.

develop in a relatively precise context the Peircean maxims honestly to seek the truth and never to block the way of inquiry.[33] Suppose that a presupposition of test *T* turned out not to be subject to an N-P test. In this case the acceptability of *H* would depend, if only indirectly, on the truth of *P*, but one would have no means for empirically testing *P*. One avenue for further inquiry into the acceptance of *H* would be blocked. Indeed, one could claim that the degree to which the acceptance of *H* is an empirical question within the N-P framework is roughly inversely proportional to the relative number of ultimate presuppositions of *T* not testable by N-P methods. If it happened that no *immediate* presupposition of *T* were N-P testable, the acceptance of *H* would surely lack any kind of "ultimate" empirical justification.

These considerations suggest the following condition for the acceptance of *H* to be ultimately justified:

(2) No ultimate presupposition of *T* may be logically impossible to test by N-P methods.

Note that this rule covers only presuppositions that might be *logically* beyond N-P methods. The satisfaction of this rule can therefore be determined by logical considerations alone, thus meeting the Humean requirement.

That N-P methods satisfy rule (2) follows from the fact, proved in section 4, that the N-P logic is presuppositionally closed. The preceding discussion suggests that any "fundamental" inductive method should exhibit this characteristic. Moreover, presuppositional closure seems to be a prerequisite for any adequate methodological account of ultimate justification. Indeed, without presuppositional closure it is hard to see how the methodologist can honestly refuse the foundationist's demand for a primary inductive method.

The concept of presuppositional closure may be further clarified by noting whether or not it holds for several other common characterizations of inductive inference. Simple enumeration may be said to be trivially closed since applications of this method presumably require no presuppositions whatsoever. The subjective Bayesian approach would seem to be presuppositionally closed if one takes prior probabilities to

[33] These maxims are present throughout Peirce's methodological writings. For an explicit statement and discussion of the rule "Do not block the way of inquiry" see (1934–58, vol. 1, bk. I, ch. 3; especially 1.135).

be presuppositions. Any prior probability distribution could be a posterior distribution relative to some other prior and some body of evidence. Induction by elimination also seems to be closed since the presupposition that the true hypothesis is a member of a given set of possibilities may be reached by elimination from a larger set of presupposed alternatives. On the other hand, purely logical accounts of inductive reasoning, e.g., Carnap's, are not closed since every system of logical probability presupposes a measure function, but no system assigns probabilities to measure functions themselves. This result might not be undesirable if one could show that the measure function is *not* an *empirical* presupposition, but I doubt this can be shown. Finally, it seems to be true generally of methods incorporating global presuppositions, e.g., measure functions, the uniformity of nature, limited independent variety, etc., that they preclude presuppositional closure. If every inductive inference must presuppose the same global principle, then the method could be closed only by begging the question.

This brief survey of inductive methods illustrates a further point concerning presuppositional closure, namely, that it can hold only for methods whose presuppositions are 'local' rather than 'global'. (The converse is not true.) Thus a necessary precondition for a successful methodological account of justification is that one's characterization require only local presuppositions. I will not attempt a precise definition of a local or a global statement since it is clear that any statistical hypothesis is local in the intended sense. Roughly speaking, global presuppositions apply to the whole universe at all times and are presupposed in every application of the methods in question.[34]

Turning to the next condition, it is useful to recall earlier philosophical debates over whether empirical meaningfulness requires only the logical possibility of verification, or whether it requires physical and perhaps even technical possibility as well. The same sort of issue arises here, though the problem is one of justification rather than meaning. Suppose at some time, *t*, an ultimate presupposition, *P*, of *T* is logically N-P testable, i.e., *P* is a statistical hypothesis, but *P* cannot be tested at that time for one of the following reasons: (a) Any N-P test would

[34] Levi (1967) explicitly distinguishes "local" from "global" justification. This type of distinction is also central to Shimony's views as expressed in his rich and inspiring paper "Scientific Inference" (1970). Shimony, who is also strongly influenced by Peirce, builds his views on a type of subjective Bayesian logic which he calls "tempered personalism."

violate a physical law; (b) Any test is beyond currently available technology; (c) Any test would violate the morals or laws of the surrounding society.[35] Would one, at *t*, be fully justified in accepting *H*?

The argument against saying the acceptance of *H* is fully justified is an extension of that given for condition (2). If *P* cannot be tested at *t*, for whatever reason, then inquiry is blocked. One who wishes to question the acceptance (or rejection) of *H* by questioning the truth of *P* cannot at that time perform an empirical test on *P*. He can only be offered the question-begging assertion that *P* would be accepted if it could be tested or will be accepted when it is eventually tested.

The argument against requiring the physical, technical, or social possibility of testing *P* is that it relativizes ultimate justification to a body of physical laws, techniques, or social customs. The first is no doubt easier to swallow than the latter two. Is it right that the acceptance of *H* should be unjustified now and justified later solely because the invention of some gadget makes it technically possible to test *P*? Again, should accepting *H* be justified in one society and not in another when everything, including all evidence, is the same in both, except for the fact that the one society permits euthanasia, thus making possible certain experiments on humans not possible in the other society? I raise these questions here mainly to make clear what is involved in granting such relativization. The amount of relativization necessary or desirable in a methodological account of ultimate justification can be better judged when we have a full set of conditions before us.

The attempt to demand more than the mere logical possibility of N-P testing ultimate presuppositions leads immediately to a further epistemic relativization. The statement that testing *P* is possible physically, technically, and perhaps even socially could only be asserted on inductive grounds. Thus any rule demanding such possibilities violates the general Humean requirement imposed above. An obvious way to resolve this conflict is to demand not that the possibilities obtain, but only that the *impossibility* of testing not be *known*. More precisely:

> (3) No ultimate presupposition of *T* may be accepted (at *t*) as being physically, technically, or socially impossible to test by N-P methods.

[35] Examples of cases (b) and (c) are easily imagined. An example of (a) might occur in astronomy if obtaining a proper sample were to require signals traveling faster than light.

Of course accepting something as physically impossible requires an inductive logic which goes beyond N-P testing, but rule (3) does not necessarily commit us even to the existence of such a method, let alone to providing a detailed characterization of its logic. And surely (3) can in principle be applied without using any inductive methods whatsoever.

Note that rule (3) introduces a whiff of the venerable but generally disreputable doctrine that knowledge may be derived from ignorance. For according to (3), the ultimate acceptability of *H* depends on our not knowing that *P* is now physically, technically, or socially impossible to test. However, while I would strenuously object to any inductive logic which, unlike the N-P theory, had a principle of indifference built into its internal characterization, it is not obvious that some appeal to ignorance is undesirable, or even avoidable, at the level of ultimate justification. Indeed it seems entirely natural that at some point in the quest for further justification one should simply reply that we have no reason for thinking otherwise. Rule (3) tells us precisely when and how such a response becomes appropriate.[36]

It is time we faced squarely the problem of stopping the Humean regress implicit in the epistemological structure of N-P tests. Since every test presupposes some nonvacuous statistical hypotheses and every statistical hypothesis might be tested, how do we justify selecting certain untested presuppositions as ultimate? According to the pragmatist tradition, the motive force of inquiry is active doubt. There is no inquiry without initial doubt, and inquiry ends when doubt is resolved. This suggests taking as ultimate presuppositions only statistical hypotheses whose truth is currently not seriously questioned.[37] Several points of clarification are necessary before this suggestion can be turned into an explicit rule.

[36] Kyburg (1961), for example, builds a principle of indifference into the characterization of his inductive logic by *defining* a random sample as one *not known* to be unrepresentative of the population. It seems to me that an adequate inductive logic will have to leave random sampling as a special empirical presupposition, always open to further investigation. Relativization to current knowledge should come later — at the level of *ultimate* justification.

[37] Recall that doubt plays a key role in Peirce's and Dewey's theory of inquiry. My notion of "seriously questioning" a hypothesis is very similar to Shimony's (1970) conception of a "seriously proposed" hypothesis, and much of his discussion is relevant here. In particular, both notions lead to Peirce's view that scientific inquiry is an essentially social phenomenon.

First, it must be emphasized that *P*'s not being actively doubted (or seriously questioned) is nothing more than a brute singular fact. The Humean regress is thus blocked by simple facts. Any attempt to make active doubt more than a mere fact opens the door to the familiar regress. For example, it might be argued that it is not enough for *P* not to be doubted — there should be *no reason* to doubt *P*. On one analysis, this formulation is unobjectionable. *P* is an empirical statistical hypothesis, and the only feature of *P* that concerns us here is its truth or falsity. Moreover, in the present discussion of ultimate justification, we may assume that the only reason for taking *P* to be true (or false) would be the outcome of an N-P test. But since ultimate presuppositions are by definition untested, it follows that we have "no reason" to doubt the truth of *P*. On the other hand, one might have intended the stronger claim that there should be reason not to doubt *P*. This interpretation, however, must be rejected because a reason not to doubt *P* would, in the present context, have to be a favorable N-P test. This way lies the regress.

A second clarification is that on my understanding of the pragmatist tradition, doubting or questioning is essentially a social phenomenon. This means that any particular ultimate justification must be relativized to some sort of scientific community, say the members of a subfield like solid state physics or learning theory. Such relativization raises a number of difficult questions. For example, what defines the relevant scientific community for test *T*? This is especially difficult since undoubtedly one of the criteria for admission will be a body of shared assumptions. If the methodological approach to ultimate justification is to be further developed, such questions will eventually have to be answered. For the moment, however, I will rest content with the following rule:

> (4) No ultimate presupposition of *T* may in fact be seriously questioned (at *t*) by members of the relevant scientific community.

The problem of providing a further rationale for this rule itself will come up again in the following section.

According to rule (4), no presupposition of *T* *is* currently under serious question. To keep faith with the methodological tradition of Peirce and Dewey, however, we should perhaps emphasize that any presupposition *can* be questioned. One may contend that this latter rule follows from rules (2) and (3) which guarantee that, to the best of our knowledge, any presupposition can be *tested*. But even if it does follow, it is worth stating separately:

(5) No ultimate presupposition of *T* may be accepted (at *t*) as being physically, technically, or socially impossible to question ("seriously").

This rule has been relativized to accepted knowledge to avoid a possible violation of the Humean requirement. Also it may turn out that the class of statistical hypotheses which are physically or technically incapable of being questioned is null. This depends on one's understanding of what it is for a member of a scientific community seriously to question a hypothesis. That a hypothesis may be socially immune to serious question seems fairly clear once it is granted that serious questioning is essentially a social activity. It would suffice, for example, that anyone attempting to voice doubts about a particular hypothesis be subject to immediate arrest and imprisonment.

One final condition seems necessary to eliminate the possibility that rule (4) is satisfied merely as a result of sheer indolence or lack of interest in finding the truth. Suppose, for example, that no one seriously questions the ultimate presuppositions of *T* because no one is really interested in the truth or falsity of *H*, or perhaps any statistical hypothesis, and therefore no one is concerned with the truth or falsity of *T*'s presuppositions. In such a situation the acceptance of *H* as a result of *T* could hardly be said to be justified in any philosophically satisfactory sense. Unless it is a general policy of the community to consider, question, and sometimes to test presuppositions, the unquestioned status of *T*'s ultimate presuppositions is irrelevant to any ultimate justification for the acceptance of *H*. This point is covered by the following suggested rule:

(6) The scientific community relevant to test *T* must be actively engaged in the pursuit of truth.

Here it is presumed that active pursuit of truth implies examining, questioning, and sometimes even testing presuppositions of statistical tests.

Rule (6) is again reminiscent of Peirce who sometimes writes as if a sincere desire to discover the truth were both necessary and sufficient for success, if only the object of the desire is pursued long enough.[38] Rule (6) also resembles Popper's claim that corroboration of a theory requires a "sincere attempt" to refute it. Popper's position, however, is complicated by the fact that while the notion of a sincere attempt at refutation is part of his characterization of corroboration, he insists that it cannot be formalized. Both opponents and supporters of Popper's position have found his views unsatisfactory and have attempted to demystify the notion of sincerely attempting to refute a theory.[39] Now while I would agree with the opinion that such a notion has no place in the *characterization* of any inductive logic, I disagree that it should be banished altogether. It seems entirely appropriate that such a notion should appear in a methodological account of ultimate *justification*. But since Popper will have no part of either induction or justification, I could hardly expect his endorsement of my program.

There is, finally, an important qualification that should be made explicit. Suppose *H* has been subjected to test *T* with ultimate presupposition *P* which is then at some later time seriously questioned. The difficulty here is that *P* refers to a series of trials that are now over, and the data that were originally sufficient for a good test of *H* assuming *P* would not in general have been sufficient for a good test of *P* as well. Any test of *P*, therefore, will require a new set of trials of an appropriate CSU. Moreover, to apply the results of this new test to the earlier experimental situation requires the assumption that the two contexts are similar in all relevant respects. But at any later time it is at least physically impossible to subject this statistical hypothesis to any N-P test. The hypothesis may, of course, be supported by a well-confirmed causal theory, but the same problem arises regarding the statistical tests that support the theory. Thus, if we are to use a methodological approach to ultimate justification in eliminating Humean skepticism, we must narrow our understanding of this approach.

There are two possibilities. One is to restrict the sense of the expres-

[38] In some passages (e.g., 1934–58, 2.709, 2.729) Peirce anticipates Reichenbach's pragmatic vindication of induction. The difficulty is that while Peirce wanted to apply these ideas to all forms of inductive reasoning, the argument holds only for the inference to a limiting relative frequency. This gap in Peirce's thought is discussed by Laudan (1973).

[39] For Popper's claim see (1959, p. 418). Kyburg's comment (1964, p. 269) that "all this psychologism is distasteful to the logician" expresses a widespread viewpoint. For "demystifications" see Salmon (1966) (anti-Popper) and Watkins (1964) (pro-Popper).

sion "can be tested" so that it means only "could have been tested at the time of the original experiment." But this restriction effectively eliminates the idea that a methodological justification is concerned with the *process* of inquiry. The second possibility is to suppose that the whole presuppositional structure of a given test is retested, with appropriate changes, whenever the acceptability of a questioned presupposition cannot be decided on the original data. In practice, of course, this is not done because we rely on causal theories to relate present and past data. But the fact that such retesting is always possible suffices to give substance to the concern with the process of inquiry that underlies the methodological approach to ultimate justification.

Ideally, fulfillment of the preceding conditions would be necessary and sufficient for the acceptance or rejection of H by test T to be justified to the extent that acceptance (or rejection) of any statistical hypothesis can be justified. Yet even a believer in the methodological program may doubt that this particular set of rules is ideal. The necessity for eliminating social blocks to the testing of presuppositions is especially controversial. There are also additional rules one might require. For example, one might wish explicitly to stipulate that no hypothesis, including H, appear more than once in any branch of a root structure. (There is obviously no reason to prevent the same presupposition from appearing in different branches.) However, repeating branches, e.g., H, T, P, T', H, T'', P, . . . , may be ruled out on the understanding that to test a hypothesis is automatically to question it, and ultimate presuppositions must, by rule 4, be unquestioned. I will not, however, pursue such matters further. Enough has been said to make clear the kind of commitments involved in a methodological approach to ultimate justification. A more pressing issue is to discover how one could show that anything like the preceding rules provide a *sufficient* account of ultimate justification in the context of statistical hypothesis testing.

8. Justification and Adequacy

How might one argue that satisfaction of the suggested methodological rules constitutes a sufficient ultimate justification for the acceptance (or rejection) of a statistical hypotheses? There seem to be only two general possibilities. One is to argue that satisfaction of the rules contributes to the goals of statistical inference as given by the N-P logic, i.e., the acceptance of true hypotheses and the rejection of false ones. The second is to claim that the rules explicate at least part of the meaning of ultimate justification or rational inference. Let us consider the first possibility first.

It is immediately obvious that there is no *logical* connection between satisfaction of the rules and either the truth of an accepted hypothesis or the correctness of the calculated error probabilities. The diligent seeker after truth may never accept a true hypothesis while the despot invariably decrees the truth on advice from his astrologer. In short, no methodological rules are logically necessary or sufficient for success now or in the indefinite future. Thus, if there is a connection between the rules for ultimate justifiability and the actual acceptance of true hypotheses, it could only be established inductively. But if N-P tests constitute our most basic inductive method, then any attempt to justify the conditions inductively introduces circularity at the meta-level. Consider an N-P test of the claim that an accepted test hypothesis is more likely to be true when the suggested conditions for ultimate justification are satisfied than when they are not. As evidence one would need cases in which the conditions were satisfied and the test hypothesis is true. But determining the truth of the test hypothesis would require an N-P test. Thus we cannot even obtain data necessary for the meta-test without assuming the ultimate justifiability of N-P testing.

In spite of this very Humean argument, it is difficult not to believe that an accepted hypothesis would more often be true when the methodological rules are satisfied than when they are not. In particular, an untested hypothesis that is nevertheless doubted by the relevant scientific community would more often be false than one that is not doubted. We know, inductively of course, that a trained intuition will pick up clues that would not count as evidence in any inductive logic. But the Humean argument does not show that such connections do not exist. It only shows that they may not be used to support the kind of justification program developed in the previous section.[40] This leaves us with the second alternative.

Philosophers of the so-called "ordinary language" school, as well as

[40] Shimony (1970) follows Peirce in appealing to natural selection in his justification of certain inductive procedures. I agree that some intuitions reinforced through natural selection play a role in scientific inquiry, though their applicability to microphysics or molecular biology seems farfetched. But the attempt to use such considerations in an account of ultimate justifications still seems to me objectionably circular.

Carnap in his latest writings,[41] have maintained that certain inductive methods are ultimately justified because these methods determine what we mean by justified inductive inference. To ask whether certain inferences are justified, therefore, is to ask a meaningless question. Applying this general approach to the present inquiry, it seems farfetched to claim that either the N-P logic of testing or the suggested methodological rules constitute an explication of what we mean by justified inductive reasoning. On the other hand, suppose we had a comprehensive inductive logic, including a precise characterization of theory testing, an interpretation of probability statements, and methodological rules like those given above. And suppose this overall account of scientific reasoning were judged superior to all other proposed systems of similar scope and precision. In this case we might well say that this system expresses our present conception of scientific reasoning, and the included schema of justification what we mean by justification in any relevant sense. On this view the problem of justification ultimately becomes part of the more general problem of judging one overall system of induction reasoning more *adequate* than another. How can we make these more general judgments?

At this point one might begin a search for still more general "criteria of adequacy" for complete systems of scientific inference. But this is a fruitless quest. The desired criteria cannot be purely logical principles, nor can they be supported inductively without circularity. Of course *after* one system has been judged superior to other proposed systems, then we might be able to formulate principles reflecting our judgments of relative adequacy, and these might help us better to understand our decisions. But such principles could not be taken as providing rational grounds for our decisions. Such decisions would have no rational grounds at all. They would merely be a causal result of our collective experience in working with the various systems in a variety of actual research contexts. Once the rival systems have been formulated, the only rational control we have on the judgment of adequacy is through the choice of contexts in which applications are attempted. Since we seek as comprehensive an overall system as possible, it seems clear we should try any proposed system in as wide a variety of scientific contexts as is practicable. But this is the most we can do.

[41] This, at least, is my understanding of the discussion in Carnap (1963).

Unfortunately this resolution for the problem of adequacy does not guarantee that any system of inductive reasoning will ever enjoy the kind of widespread approval enjoyed by current systems of deductive reasoning. There is, however, some ground for hope in the fact that there are very few systems that are sufficiently comprehensive and that have been sufficiently well developed even to be active candidates. At the moment the only really serious alternative to something like the N-P logic of testing is a system of Bayesian inference utilizing subjective probabilities. It seems not unreasonable to hope that patient attempts to apply these rival logics in a variety of research situations will make clear which is the more adequate account of scientific inference.

There are those who will feel that by allowing judgments of adequacy to be merely causal outcomes of scientific experience, I have abandoned the justification problem, not resolved it. I would reply that my position simply reflects our true situation regarding inductive reasoning (and probably deductive too!). In this I am in general agreement with Hume for whom inductive reasoning was ultimately a matter of animal habit. This is also the general position of such modern philosophers as Nelson Goodman, who appeals to the simple facts of past linguistic usage, and W. V. O. Quine, who now grounds epistemology in empirical psychology. The main difference between these views and mine is not that one stops trying to give reasons, but where and how one stops.[42]

On my account there are actually two points at which reasons give way to brute fact. One is at the object level, so to speak: some ultimate presuppositions of any test will be unquestioned, and one need not say why. The second is at the meta-level: judgments of adequacy are not justified by reasons. But it is never the case that the acceptance of any specific test hypothesis is itself a matter of brute fact. This remains under rational, objective control, though of course relative to the adequacy of the overall system and the ultimate presuppositions of some particular statistical test. In this way it is possible to insist that inductive

[42] Just as the "new riddle of induction" is only a linguistic version of Hume's old riddle, so Goodman's (1955) entrenchment theory is just a linguistic version of Hume's theory of habits, though more precisely articulated. That Quine (1969) does not make more of the *social* basis of epistemology may be a reflection of his philosophical upbringing at a time when the logic/psychology distinction and the reducibility of sociology to psychology were taken for granted. Then again, knowing has almost always been taken to be a function of the individual "soul." In this the pragmatists are an exception.

inference is a rational process while admitting an essential justificatory role for some brute facts.

9. The Metaphors of Ultimate Justification: Foundations, Nets, and Roots

My overall program for the philosophy of inductive reasoning is far from completed. Much remains to be done in characterizing the logic of statistical hypothesis testing and the logic of theory testing. In addition, the kinds of case studies of actual research that might produce widespread agreement on the adequacy of these methods are practically nonexistent. As it is impossible to pursue these tasks here, I will content myself with some concluding remarks on the metaphor underlying the methodological justification of statistical hypothesis testing presented above. When one reaches the limits of rational discussion, in any subject, the importance of a good metaphor should not be overlooked.

Having rejected the church and reason as guarantors of knowledge, classical empiricists looked to 'experience' as the source and foundation of all knowledge. The implicit ideal was to use only concepts somehow abstracted directly from experience and to admit only hypotheses logically derivable from statements irrefutably certified by experience alone. Knowledge was thus pictured as a solid building rising on unshakable foundations. This metaphor proved remarkably durable. Sustained by Mach's phenomenalism and the logical constructions of Russell's *Principia*, the foundationist picture, in the form of Carnap's *Aufbau*, inspired the Viennese branch of empiricism. Yet even the logical empiricists soon learned that this classical ideal is unattainable. But what is to replace it?

Here we are not concerned with how one obtains a suitable store of singular observation statements, but only with how these may be used to substantiate singular predictions or general laws, especially statistical laws. Once it is admitted that this cannot be done with standard deductive tools alone, there are two major strategies that may be pursued. One is to use the observation statements to assign (inductive) probabilities to all other statements. If, as in Carnap's program, this can be done with a purely semantical probability relation, much of the spirit of the original foundationist program is preserved. If one has doubts about using logical probabilities, a retreat to subjective probabilities is always possible.[43] The second major approach is to construct an inductive logic which contains a bona fide inductive rule of inference, i.e., one that validates empirical statements themselves and not merely probability relations with observation statements. Here one retains the spirit of the original program by insisting that the inductive rule of inference be a *primary* rule, i.e., one whose application presupposes only observation statements. Reichenbach's rule of induction is the best known rule of this type.

These weakened versions of foundationism still have many supporters, both explicit and tacit. Yet there are also many who would like to abandon foundationism altogether. For the latter, Quine has provided the tempting metaphor of knowledge as a "net" or, more recently, as a "web of belief." [44] One obvious possibility, therefore, would be to elucidate my conditions for methodological justification in terms of Quine's net metaphor. This move would be easy because Quine and his followers have never developed a precise characterization of statistical inference. The N-P logic of testing might therefore be viewed as giving the detailed structure of parts of the Quinean net. Tempting as this move might be, it seems to me mistaken.

Quine's net is really Neurath's ship.[45] The elements, i.e., the sea and wind, are experience. Owing to the impact of experience and the lack of a dry dock, the inhabitants of the ship must continually rebuild their craft while at sea. The only necessity is keeping out the wind and the sea. Beyond this the primary principles of shipbuilding are simplicity and elegance, however these may be defined.

My first objection to Quine's picture is merely the intuition that knowledge should be more than something bobbing about on the surface of experience, no matter how internally elegant that something might be. Knowledge should somehow reach into experience. At the very least, Neurath's ship could use an anchor. But perhaps this intuition is merely the product of nostalgia for the foundationist picture.

[43] Jeffrey (1965) is a good example of someone who has made this shift.

[44] The classic source of the net metaphor is Quine's "Two Dogmas of Empiricism" reprinted in Quine (1953) and elsewhere. *The Web of Belief* (1970) is a recent elementary text by Quine and Ullian.

[45] Recall Otto Neurath's famous remark quoted by Quine at the beginning of *Word and Object* (1960, p. viii): "Wie Schiffer sind wir, die ihr Schiff auf offener See umbauen müssen, ohne es jemals in einen Dock zerlegen und aus besten Bestandteilen neu errichten zu können."

My second objection is that the net metaphor is too vague, even for a metaphor. While it is compatible with a methodological justification for N-P testing, it is also compatible with a Carnap-style foundationism. One need only interpret the strings of the net as logical or subjective probability relations. This very natural interpretation virtually eliminates any substantial differences between Quine and Carnap.

The appropriate metaphor is suggested by the epistemological structure of N-P tests — a system of roots. This suggests the organic metaphor of knowledge as a tree with branches growing up and roots growing down into the soil of experience. The roots, however, do not consist solely of observations, but, like the branches, contain statistical hypotheses as well. Moreover, there is no theoretical limit to the size of the tree so long as the roots are extensive enough and deep enough. The principal value of this metaphor is that it makes intuitively clear that there can be an extensive structure of knowledge rising above the ground of experience without there being any foundation at all. Roots are sufficient. In addition, a tree metaphor goes well with a methodological account of ultimate justification. The methodological rules give conditions on the state of the roots and how they must grow if we are to have a genuine "tree of knowledge." [46]

The picture of knowledge as a plant with roots and branches is in keeping with the contemporary passion for things organic. And a pragmatic, methodological account of ultimate justification reflects another current in contemporary thought, an interest in the social aspects of all things. Whether these themes will prove lasting enough to deserve a place in our epistemology is now impossible to judge. If my conception of judgments of adequacy is correct, our task is to develop comprehensive systems and try them in scientific contexts. Only in this way can we decide whether, having abandoned the foundationist picture, knowledge is now to be viewed as a net, a tree, or something completely different.

[46] There is an obvious parallel between a tree of knowledge supported by roots and Popper's metaphor of a building supported by piles driven into the swamp of experience (1959, p. 111). The structure of the N-P inductive logic makes the root metaphor clearly preferable.

REFERENCES

Anscombe, F. J. (1963). "Tests of Goodness of Fit," *Journal of the Royal Statistical Society B*, vol. 25, pp. 81–94.

Arbuthnot, J. (1710). "An Argument for Divine Providence Taken from the Constant Regularity Observed in the Births of Both Sexes," *Philosophical Transactions of the Royal Society of London*, vol. 27, pp. 186–90.

Bar-Hillel (1968). "The Acceptance Syndrome," in I. Lakatos, ed., *The Problem of Inductive Logic*. Amsterdam: North Holland.

Birnbaum, A. (1962). "On the Foundations of Statistical Inference," *Journal of the American Statistics Association*, vol. 57, pp. 269–306.

Birnbaum, A. (1969). "Concepts of Statistical Evidence," in S. Morgenbesser, P. Suppes, and M. White, eds., *Philosophy, Science and Method*. New York: St. Martin's.

Birnbaum, A. (1971). "A Perspective for Strengthening Scholarship in Statistics," *The American Statistician*, vol. 25, no. 3, pp. 14–17.

Black, M. (1971). *Margins of Precision*. Ithaca, N.Y.: Cornell University Press.

Carnap, R. (1950). *Logical Foundations of Probability*. 2nd ed. Chicago: University of Chicago Press, 1962.

Carnap, R. (1963). "Replies and Systematic Expositions," in P. A. Schilpp, ed., *The Philosophy of Rudolf Carnap*. La Salle, Ill.: Open Court.

de Finetti, B. (1937). "Foresight: Its Logical Laws, Its Subjective Sources," in H. E. Kyburg and H. E. Smokler, eds., *Studies in Subjective Probability*. New York: John Wiley, 1964.

Dewey, J. (1939). *Logic: The Theory of Inquiry*. New York: Holt, Rinehart, and Winston.

Earman, J. (1971). "Laplacian Determinism," *Journal of Philosophy*, vol. 68, pp. 729–44.

Edwards, A. W. F. (1971). *Likelihood*. Cambridge: At the University Press.

Feyerabend, P. K. (1962). "Explanation, Reduction and Empiricism," in H. Feigl and G. Maxwell, eds., *Minnesota Studies in the Philosophy of Science*, vol. 3. Minneapolis: University of Minnesota Press. Pp. 28–97.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. London: Oliver and Boyd.

Fisher, R. A. (1935). *The Design of Experiments*. London: Oliver and Boyd.

Fisher, R. A. (1950). *Contributions to Mathematical Statistics*. New York: John Wiley.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. New York: Hafner.

Giere, R. N. (1973a). "Objective Single Case Probabilities and the Foundations of Statistics," in P. Suppes, L. Henkin, A. Joja, and C. R. Moisil, eds., Logic, *Methodology and Philosophy of Science*, vol. 4: *Proceedings of the 1971 International Congress, Bucharest*. Amsterdam: North Holland.

Giere, R. N. (1973b). "History and Philosophy of Science: Intimate Relationship or Marriage of Convenience?" *British Journal for the Philosophy of Science*, vol. 24, pp. 282–97.

Giere, R. N. (1973c). Review of D. H. Mellor, *The Matter of Chance*, in *Ratio*, vol. 15, pp. 149–55.

Giere, R. N. (1975). "Empirical Probability, Objective Statistical Methods, and Scientific Inquiry," in C. Hooker and W. Harper, eds., *Foundations of Probability and Statistics and Statistical Theories in Science*. Dordrecht: Reidel.

Gilles, D. (1973). An *Objective Theory of Probability*. London: Menthuen.

Glymour, C. (1971). "Determinism, Ignorance, and Quantum Mechanics," *Journal of Philosophy*, vol. 68, pp. 744–51.

Good, I. J. (1950). *Probability and the Weighing of Evidence.* London: Griffin.

Good, I. J. (1965). *The Estimation of Probabilities.* Cambridge, Mass.: M.I.T. Press.

Goodman, N. (1955). *Fact, Fiction and Forecast.* Cambridge, Mass.: Harvard University Press. 2nd ed., 1965.

Hacking, I. (1965). *Logic of Statistical Inference.* Cambridge: At the University Press.

Hanson, N. R. (1958). *Patterns of Discovery.* Cambridge: At the University Press.

Hays, W. L. (1963). *Statistics for Psychologists.* New York: Holt, Rinehart, and Winston.

Hintikka, J. (1966). "A Two-Dimensional Continuum of Inductive Methods," in J. Hintikka and P. Suppes, eds., *Aspects of Inductive Logic.* Amsterdam: North Holland. Pp. 113–32.

Hodges, J. L., and E. L. Lehman (1964). *Basic Concepts of Probability and Statistics.* San Francisco: Holden and Day.

Jeffrey, R. C. (1956). "Valuation and Acceptance of Scientific Hypotheses," *Philosophy of Science,* vol. 23, pp. 237–46.

Jeffrey, R. C. (1965). *The Logic of Decision.* New York: McGraw-Hill.

Kempthorne, O., and Leroy Folks (1971). *Probability, Statistics, and Data Analysis.* Ames: Iowa State Press.

Kendall, M. G., and A. Stuart (1958–66). *The Advanced Theory of Statistics,* 3 vols. London: Griffin.

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

Kyburg, H. E., Jr. (1961). *Probability and the Logic of Rational Belief.* Middletown, Conn.: Wesleyan University Press.

Kyburg, H. E., Jr. (1964). "Recent Work in Inductive Logic," *American Philosophical Quarterly,* vol. 1, pp. 249–87.

Kyburg, H. E., Jr. (1968). "The Rule of Detachment in Inductive Logic," in I. Lakatos, ed., *The Problem of Inductive Logic.* Amsterdam: North Holland. Pp. 98–119.

Kyburg, H. E., Jr. (1970). "Conjunctivitis," in M. Swain, ed., *Induction, Acceptance and Rational Belief.* Dordrecht: Reidel.

Kyburg, H. E., Jr., and H. E. Smokler, eds. (1964). *Studies in Subjective Probability.* New York: John Wiley.

Laplace, P. Simon Marquis de (1814). *A Philosophical Essay on Probabilities.* Paris.

Laudan, L. (1973). "Peirce and the Trivialization of the Self-Correcting Thesis," in R. N. Giere and R. S. Westfall, eds., *Foundations of Scientific Method: The Nineteenth Century.* Bloomington: Indiana University Press.

Lehman, E. L. (1959). *Testing Statistical Hypotheses.* New York: John Wiley.

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

Lindley, D. V. (1971). *Bayesian Statistics, a Review.* Philadelphia, Pa.: Society for Industrial and Applied Mathematics.

Luce, R. D., and H. Raiffa (1957). *Games and Decisions.* New York: John Wiley.

Mellor, D. H. (1971). *The Matter of Chance.* Cambridge: At the University Press.

Neyman, J. (1950). *First Course in Probability and Statistics.* New York: Henry Holt.

Neyman, J. (1957). "'Inductive Behavior' as a Basic Concept of Philosophy of Science," *Review of the International Statistical Institute,* vol. 25, pp. 7–22.

Neyman, J. (1967). *A Selection of Early Statistical Papers of J. Neyman.* Berkeley: University of California Press.

Neyman, J., and E. S. Pearson (1967). *Joint Statistical Papers.* Berkeley: University of California Press.

Nicod, J. (1930). *Foundations of Geometry and Induction*, trans. P. P. Wiener. London: Routledge and Kegan Paul. Trans. John Bell and Michael Woods, London, 1969.

Pap, A. (1962). *An Introduction to the Philosophy of Science*. New York: Free Press.

Pearson, E. S. (1966). "The Neyman-Pearson Story: 1926–34," in F. N. David, ed., *Research Papers in Statistics*. New York: John Wiley.

Peirce, C. S. (1934–58). *Collected Papers of Charles Sanders Peirce*, 8 vols. Cambridge, Mass.: Harvard University Press.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

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

Quine, W. V. O. (1953). *From a Logical Point of View*. Cambridge, Mass.: Harvard University Press.

Quine, W. V. O. (1960). *Word and Object*. Cambridge, Mass.: M.I.T. Press.

Quine, W. V. O. (1969). *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Quine, W. V. O., and J. S. Ullian (1970). *The Web of Belief*. New York: Random House.

Reichenbach, H. (1949). *The Theory of Probability*. Berkeley: University of California Press.

Rogers, B. (1971). "Material Conditions on Tests of Statistical Hypotheses," in R. C. Buck and R. Cohen, eds., *Boston Studies in the Philosophy of Science*. New York: Humanities Press.

Salmon, W. C. (1966). *The Foundations of Scientific Inference*. Pittsburgh: University of Pittsburgh Press.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: John Wiley.

Savage, L. J. (1962). "Bayesian Statistics," in R. E. Machol and P. Gray, eds., *Recent Developments in Decision and Information Processes*. New York: Macmillan.

Savage, L. J. (1967). "Implications of Personal Probability for Induction," *Journal of Philosophy*, vol. 64, pp. 593–607.

Shimony, A. (1970). "Scientific Inference," in R. G. Colodny, ed., *The Nature and Function of Scientific Theories*. Pittsburgh: University of Pittsburgh Press.

Strawson, P. F. (1952). *Introduction to Logical Theory*. New York: John Wiley.

Suppes, P. (1962). "Models of Data," in E. Nagel, P. Suppes, and A. Tarski, eds., *Logic, Methodology and Philosophy of Science*. Stanford: Stanford University Press.

Tukey, J. W. (1961). "The Future of Data Analysis," *Annals of Mathematical Statistics*, vol. 33, pp. 1–67.

van Fraassen, B. C. (1971). "On the Extension of Beth's Semantics of Physical Theories," *Philosophy of Science*, vol. 38, pp. 325–39.

van Fraassen, B. C. (1972). "A Formal Approach to the Philosophy of Science," in R. G. Colodny, ed., *Paradigms and Paradoxes*. Pittsburgh: University of Pittsburgh Press. Pp. 303–66.

von Mises, R. (1957). *Probability, Statistics and Truth*. London: Allen and Unwin.

Wald, A. (1950). *Statistical Decision Functions*. New York: John Wiley.

Watkins, J. (1964). "Confirmation, the Paradoxes and Positivism," in Mario Bunge, ed., *The Critical Approach to Science and Philosophy*. New York: Free Press.