that they have been putting taxa names in a metaphysical category to which such names do not belong. Instead of being highly aberrant classes, they are typical individuals (see Individuality).

But this discussion does not entail anything about the metaphysical nature of the species category itself. Species taxa are spatiotemporally restricted; the species category is not. It has all the generality needed to count as a kind. Species as evolvers are not restricted to Earth. In all probability, they have evolved numerous times throughout the universe. A particular lineage cannot evolve more than once, but lineages as such can recur. In addition, if species are that which evolves, then they function in an important scientific theory. The net effect is that species as such are a natural kind. *Homo sapiens* as a taxon is not a natural kind; the species category is.

DAVID HULL

References

Claridge, M. F., H. A. Dawah, and M. R. Wilson (eds.) (1997), *Species: The Units of Biodiversity*. London: Chapman and Hall.

Cracraft, J. (1983), "Species Concepts and Speciation Analysis," in *Current Ornithology*, vol. 1. New York: Plenum Press, 159–187.

de Queiroz, K. (1998), "The General Lineage Concept of Species, Species Criteria, and the Process of Speciation," in D. J. Howard and S. H. Berlocher (eds.), *Endless Forms: Species and Speciation*. Oxford: Oxford University Press, 57–75.

——— (1999), "The General Lineage Concept of Species and the Defining Properties of the Species Category," in R. A. Wilson (ed.), *Species: New Interdisciplinary Essays*. Cambridge, MA: MIT Press, 49–89.

Donoghue, M. J. (1985), "A Critique of the BSC and Recommendations for a Phylogenetic Alternative," *Bryologist* 83: 172–181.

Eldredge, N., and J. Cracraft. (1983), "Species Concepts and Speciation Analysis," *Current Ornithology* 1, 159–187.

Ereshefsky, M. (1992), *The Units of Evolution: Essays on the Nature of Species*. Cambridge, MA: MIT Press.

Ghiselin, M. T. (1974), "A Radical Solution to the Species Problem," *Systematic Zoology* 25: pp. 536–544.

Hennig, W. (1966), *Phylogenetic Systematics*. Chicago, IL: University of Illinois Press.

Hull, D. L. (1976), "Are Species Really Individuals?" *Systematic Zoology* 25: 174–191.

Kitcher, P. (1984), "Species," *Philosophy of Science* 51: 308–333.

Mayden, R. L. (1997), "A Hierarchy of Species Concepts: The Denouement in the Saga of the Species Problems," in M. F. Claridge, H. A. Dawah, and M. R. Wilson (eds.), *Species: The Units of Biodiversity*. London: Chapman and Hall, 381–424.

Mayr, E. ([1942] 1964), *Systematics and the Origin of Species: From the Viewpoint of a Zoologist*, 2nd ed. New York: Dover.

——— (1969), *Principles of Systematic Zoology*. New York: McGraw-Hill.

Mishler, B. D. (1985), "The Morphological, Developmental, and Phylogenetic Basis of Species Concepts in Bryophytes," *Bryologist* 88: 207–214.

Mishler, B. D., and R. N. Brandon. (1987), "Individualism, Pluralism, and the Phylogenetic Species Concept," *Biology and Philosophy* 2: 397–414.

Nixon, K. C., and Q. D. Wheeler. (1990), "An Amplification of the Phylogenetic Species Concept," *Cladistics* 6: 211–223.

Otte, D., and J. A. Endler (eds.) (1989), *Speciation and Its Consequences*. Sunderland, MA: Sinauer.

Rosen, D. E. (1979), "Fishes from the Uplands and Intermontane Basins of Guatemala: Revisionary Studies and Comparative Geography," *Bulletin of the American Museum of Natural History* 162: 267–376.

Simpson, G. G. (1961), *Principles of Animal Taxonomy*. New York: Columbia University Press.

Sneath, P. H. A., and R. R. Sokal. (1973), *Numerical Taxonomy*. San Francisco: Freeman.

Templeton, A. R. (1989), "The Meaning of Species and Speciation," in D. Otte and J. A. Endler (eds.), *Speciation and Its Consequences*. Sunderland, MA: Sinauer, 3–27.

Wiley, E. O. (1981), *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*. New York: John Wiley.

Wilson, R. A. (1999), *Species: New Interdisciplinary Essays*. Cambridge, MA: MIT Press.

See also **Conservation Biology; Evolution; Individuality; Natural Selection**

PHILOSOPHY OF STATISTICS

Philosophy of statistics may be seen to encompass the epistemological, conceptual and logical problems revolving around the use and interpretation of the methods of mathematical statistics. In contrast to the better known philosophies of science, physics, and mathematics, work in philosophy of

statistics is as likely to be engaged in by practicing statisticians as by philosophers of science. Accordingly, contributions to philosophy of statistics might be regarded just as much as contributions to statistics as to philosophy of science. To make this entry useful and of manageable length, it focuses on the main philosophical debates relating to the modern methodology for *statistical inference*: significance tests, hypothesis testing, confidence interval estimation, likelihood, and Bayesian methods. This still leaves a huge territory marked by seventy years of debates widely known for reaching unusual heights both of passion and of technical complexity. To get a handle on the movements and cycles without too much oversimplification or distortion, three main waves of debates in philosophy of statistics will be distinguished: 1930–1960, 1960–1980, and 1980 to the present.

A core question that underlies the debates is: What is the nature and role of probabilistic concepts, methods, and models in making inferences in the face of limited data, uncertainty, and error? The different answers to this question have immediate ramifications for all of the central issues around which much of the debates revolve: what tasks do mathematical methods of statistics perform? And what criteria or principles are appropriate for evaluating them?

Two Roles for Probability in Inference

There are two distinct philosophical traditions regarding the role of probability in statistical inference in science. In one, probability is used to provide a post-data assignment of degree of probability, confirmation, support, or belief in a hypothesis, while in a second, probability is used to assess the probativeness, reliability, trustworthiness, or severity of a test or inference procedure.

Confirmation Theory

Conceding that all attempts to solve the problem of induction (see Induction, Problem of) suffered from circularity (Salmon, 1967), philosophers of induction (e.g., in the 1970s) turned their attention instead to constructing *logics of induction* or *confirmation theories* that would, ideally, reflect "inductive intuition." The goal would be to supply means to compute the degree of *evidential relationship* between given evidence statements, e, and a hypothesis, H (see Confirmation Theory). A natural place to look for such a computation is the definition of conditional probability, or *Bayes's theorem*:

$$P(H \mid e) = P(e \mid H)P(H)/P(e)$$

where $P(e) = P(e \mid H)P(H) + P(e \mid \neg H)P(\neg H)$.

Computing $P(H \mid e)$, the *posterior probability*, requires starting out with a probability assignment to all of the members of $\neg H$, and a major source of difficulty through all three waves is how to obtain, justify, and interpret these prior probabilities. Insofar as the computed degrees of confirmation are viewed as analytic and a priori, their relevance for predicting and learning about empirical phenomena is problematic; insofar as they measured subjective degrees of belief, their relevance for giving objective guarantees of reliable inference is unclear (see Bayesianism; Confirmation Theory; Inductive Logic).

The Error-Probability Philosophy ('Sampling Theory')

A distinct philosophical tradition uses probability to characterize a procedure's overall reliability in a series of (actual or hypothetical) experiments or in *repeated sampling* (hence, 'sampling theory'). These probabilistic properties of statistical procedures are called *error frequencies* or *error probabilities* (e.g., significance levels, confidence levels). Deliberately designed to reach conclusions about statistical parameters without invoking prior probabilities in hypotheses, error probabilistic methods use probability to quantify how *frequently* methods discriminate between alternative hypotheses and how *reliably* they facilitate the detection of error. As with logics of confirmation, there are connections with philosophy of induction, as in Peirce, Braithwaite, and, to some extent, Popper (see Popper, Karl Raimund). These two contrasting philosophies of the role of probability in statistical inference correspond to the core issues at the heart of the debate in all three waves of philosophy of statistics.

The First Wave

Quantitative methods of statistical inference involve drawing conclusions about parameters on the basis of the observed values of random variables. Statistical methods may be seen to connect questions about the phenomenon or data-generating source to questions about distributions of random variables that model the data-generating source or population. Thus the conception of a statistical model wherein these parameters are defined is an important component of statistical

inference methods. The area of model specification and model selection has its own set of philosophical issues that will not be taken up here. A statistical hypothesis cannot be just any claim, but must give probability assignments to the different experimental outcomes or sample space Ξ, typically in terms of the parameters of the model. That is, for any x in Ξ, H assigns "the probability of x *under* H," written $P(x;H)$. This notation helps avoid confusion between a probabilistic computation under a model and conditional probabilities needed for Bayes's theorem, $P(x \mid H)$, without prejudging issues. (An alternative notation some find useful is $P(x \mid\mid H)$; see Friedman 1995).

Fisherian "Simple" Significance Tests

The modern approach to statistical inference was initiated by Fisher, who introduced the main concepts and procedures of statistical significance tests. Fisher's strong objections to Bayesian inference (Fisher 1935, 1955), and in particular to the use of prior distributions, led Fisher to develop ways to express the uncertainty of inferences without deviating from frequentist probabilities.

The significance test is a procedure with the following components: there is a null hypothesis H_0 that is an assertion about the distribution of the sample $X = (X_1, \ldots, X_n)$, and a function of the sample, $d(X)$, the *test statistic*, which measures the difference between the data $x_0 = (x_1, \ldots, x_n)$, and null hypothesis H_0. The larger the value of $d(x_0)$, the further the outcome is from what is expected under H_0, with respect to the particular question being asked (x_0 represents a particular realization of X). For an observed difference $d(x_0)$, the test computes the p-value, or the probability of a difference larger than $d(x_0)$, computed under the assumption that H_0 is true:

$$p(x_0) = P(d(\mathrm{X}) > d(x_0); H_0).$$

The p-value may be regarded as a measure of discordancy from H_0: the smaller the significance level, the greater the discordance between x_0 and H_0 (Kempthorne and Folks 1971).

Fisher described the significance test as a procedure for rejecting the null hypothesis and inferring that the phenomenon has been "experimentally demonstrated" (Fisher 1935, 14), where the latter inference corresponds to finding a small p-value, such as .05 or .01. How to justify this is a point of philosophical debate. One highly influential example is this. Suppose that x_0 is evidence against H_0 just in case x_0 is statistically significant at a small level p (or smaller). Then p is the maximal probability of rejecting H_0 when H_0 is actually a correct description of the underlying data-generating mechanism. So there is only a small probability of erroneously rejecting H_0, i.e., committing what Neyman-and Pearson call a *type I error* (Cox, 1958). Commonly used significance tests—Pearson's chi-square goodness of fit, the Student t test, the F test in analysis of variance—are regularly used to distinguish real effects of importance from apparent effects actually due to random sampling or uncontrolled variability.

The Alternative or "Non-Null" Hypothesis

Evidence against H_0 would seem to indicate evidence for some alternative, if only for a directional departure from the null value in a given direction. Although Fisherian significance tests strictly consider only the null hypothesis, Neyman and Pearson tests introduce as well an alternative H_1. Despite the bitter disputes with Fisher that were to erupt soon after their early developments of tests, Neyman and Pearson, at the outset, regarded their work as merely placing Fisherian tests on firmer logical footing by taking explicit account of an alternative to the null hypothesis.

Neyman-Pearson (N-P) Tests The N-P hypothesis test, mathematically considered, is a rule that maps each possible outcome $x = (x_1, \ldots, x_n)$ onto one of two hypotheses, the test or null hypothesis H_0 or an alternative hypothesis H_1. As in the Fisherian (simple) significance test, there is a test statistic $d(X)$, in terms of which the test rule is defined. In the N-P test, however, the values of $d(X)$ that will be taken to reject H_0 are fixed at the outset, by a predesignated choice of significance level. Most importantly, the null and alternative of an N-P test exhaust the parameter space of the statistical model, whereas in the Fisherian test there is the single null hypothesis, as against its logical complement.

The N-P error probabilities are computed under the assumption that the statistical model is adequate; what is being tested are the values of one or more parameters governing the distribution. For simplicity, illustrations here keep to the case of only one unknown parameter. Although N-P theory provides distinct tests of the assumptions of the statistical model, and the whole issue of model validation is important philosophically, the matter will not be explicitly discussed here.

Example. Test T(α): Consider a random sample of size n, $\boldsymbol{X} = (X_1, \ldots, X_n)$, where it is assumed that each X_i is normal $N(\mu,\sigma^2)$, independent and identically distributed (IID). Test $T(\alpha)$ denotes the familiar test of H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$, where H_0 is the null, and H_1 the alternative hypothesis. Because H_1 includes only positive discrepancies from H_0, this is called a one-sided test. For simplicity, let the standard deviation σ be known—for instance, let $\sigma = 1$. The test statistic for T(α) is: $d(\mathbf{X}) = (\bar{X} - \mu)/\sigma_{\bar{X}}$, where $\bar{X}$ is the sample mean with standard deviation $\sigma_{\bar{X}} = (\sigma/\sqrt{n})$. The N-P test with *significance level* α rejects H_0 with data x_0 if and only if $d(x_0)$ reaches the preset significance level α—for instance, $c_\alpha = 1.96$ for $\alpha = .025$, so

Test T(α): if $d(\boldsymbol{x}_0) > c_\alpha$, reject H_0,
if $d(x_0) \leq c_\alpha$, accept H_0.

The set of all outcomes that lead to "reject H_0" is called the *rejection region*. "Accept" and "reject" should be regarded as parts of the mathematical apparatus whose interpretation must be separately considered.

The test is specified so that the probability of a type I error, α, is fixed at some small number, such as .05 or .01, the *significance level* of the test:

Type I error probability
$= P(\text{Test T}(\alpha) \text{ Rejects } H_0; H_0) \leq \alpha.$

Since "Test T(α) Rejects H_0" iff $\{d(X) > c_\alpha\}$, it follows that

Type I error probability $= P(d(X) > c_\alpha; H_0) \leq \alpha.$

N-P test principles then seek out the test that at the same time has a small probability of committing a type II error, β. Since the alternative hypothesis H_1, as is typical, contains more than a single value of the parameter, it is *composite*, the type II error probability is evaluated at a specific point $\mu = \mu_1$, and thus is abbreviated $\beta(\mu_1)$:

$P(\text{Test T}(\alpha) \text{ does not reject } H_0; \mu = \mu_1) = P(d(X) \leq c_\alpha; H_0) = \beta(\mu_1)$, for $\mu_1 > \mu_0$.

The "best" test with significance level β (if it exists) is the one that at the same time minimizes the value of β for all $\mu_1 > \mu_0$, or equivalently, maximizes the *power*:

$\text{POW}(\text{T}(\alpha); \mu_1) = \text{P}(d(\text{X}) > \text{c}_\alpha; \mu_1)$, for all $\mu_1 > \mu_0$.

T(α) is said to be a *uniformly most powerful* (UMP) α significance level test. Letting $\alpha = .025$, $T(\alpha)$ If $d(\text{x}) > 1.96$, reject H_0. The rejection region for the corresponding two-sided .05 test, H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$, abbreviated as T(2α) is: $\{\text{x} : |\text{d}(\text{x}_0)| > 1.96\}$.

Error Probabilities Versus Conditional Probabilities

Confusion often results from interpreting the type I error probability: $\text{P}(\text{d}(\boldsymbol{X}) > \text{c}_\alpha; \text{H}_0)$ as a conditional probability statement of the form: $P(d(\boldsymbol{X}) > c_\alpha | H_0)$. From the definition of conditional probability it follows that

$$P(d(\boldsymbol{X}) > \text{c}_\alpha | H_0) = [P(d(\boldsymbol{X}) > c_\alpha, \mu = \mu_0)]/P(\mu = \mu_0).$$

However, neither the numerator $P(d(X) > c_\alpha, \mu = \mu_0)$ nor the denominator $P(\mu = \mu_0)$ of this ratio are meaningful unless the parameter may be assumed to be a random variable, as in a Bayesian approach (see Bayesianism).

In the N-P testing paradigm, there is no probability assignment to the conjunctive event $(d(\boldsymbol{X}) > c_\alpha, \mu = \mu_0)$ or to $(\mu = \mu_0)$. The statement $P(d(X) > \text{c}_\alpha; H)$, should be interpreted as the probability of rejecting H_0 when evaluated under the hypothetical scenario that the observed outcome x_0 has arisen from the distribution described in H_0. Within the error probability (frequentist) framework, a statistical hypothesis H either does or does not adequately describe the process generating the data. There is no suggestion that any H is precisely true; indeed, the purpose of tests is to evaluate discrepancies of specified sorts. But probabilities enter in this evaluation only as error probabilities.

Inductive Behavior Philosophy

Philosophical issues and debates arise once one begins to consider the uses to which these formal statistical tools might be put, the interpretations of the formal apparatus, and the justifiability of associated principles of tests. The proof by Neyman and Pearson of the existence of best tests set the stage for the mathematical development of statistical tests as rigorous rules for "deciding" to accept or reject hypotheses. In this conception, to infer the conclusion of the significance testing argument, 'data x_0 is evidence against H_0' or 'x_0 indicates the falsity of H_0,' is to take a decision of a sort, with a calculable risk. Wishing to draw a stark contrast between this conception of tests and those of Fisher as well as Bayesians (Jeffreys), Neyman declared that the goal of tests is not to adjust beliefs but rather to "adjust behavior" to limited amounts of data. Tests, accordingly, are not rules of inductive inference but rules of behavior. The value of tests as

rules of behavior is that "it may often be proved that if we behave according to such a rule ... we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false" (Neyman and Pearson 1933, 142).

Debates Between Fisher and Neyman and Pearson: The 1950s

The dispute between "inductive behavior" and "inductive inference" coming on top of the break between Fisher and Neyman, which began in 1935, commingled philosophical, statistical, and personality clashes. Fisher (1955) denounced the way that Neyman and Pearson transformed "his" significance tests into "acceptance procedures," wherein tests are viewed as mechanical rules or recipes for deciding to accept or reject statistical hypothesis H_0, and the concern has more to do with speeding up production or making money than in learning about phenomena. In responding to Fisher, Pearson clearly distanced himself from Neyman's "inductive behavior" jargon, calling it "Professor Neyman's field rather than mine" (Pearson 1955, 207). However, Pearson protested that neither he nor Neyman were "speaking of the final acceptance or rejection of a scientific hypothesis on the basis of statistical analysis. ... Indeed, from the start we shared Professor Fisher's view that in scientific enquiry, a statistical test is 'a means of learning'" (204–205).

Neyman, too, despite promoting "inductive behavior as a major concept in philosophy of science" (1957a), clearly denounced "mechanical" uses of significance tests (in responding to Fisher), and had no hesitation in using N-P tests for "inference" or reaching "conclusions." Tracing out the thrust and parry between Neyman, Pearson, and Fisher in the 1950s will amply reward those interested in what the key players "really thought." Later on, the N-P tests became so formally entrenched in the decision-theoretic framework of Wald (1950) that many of the qualifications by Neyman and Pearson in the first wave have been overlooked in the philosophy of statistics literature.

Confidence Interval Estimation Procedures

Statistical inference can take the form of estimation procedures as well as tests. In confidence interval (CI) estimation procedures, a statistic is used to set upper or lower (one-sided) or both (two-sided) bounds. The concept of a confidence interval with a frequentist interpretation was first introduced by Neyman (1935) as a way to extend point estimation to interval estimation, with a predesignated error rate. For a parameter, say, μ, a $(1\text{-}\alpha)$ confidence interval estimation procedure leads to estimates of form:

$$\mu = \bar{X} \pm e$$

Different sample realizations x lead to different estimates, but one can ensure that $(1 - \alpha)$ 100% of the time the true parameter value μ, whatever it may be, will be included in the interval formed.

Dualities Between One- and Two-Sided Intervals and Tests

There exists a duality relationship between CIs and hypothesis tests that can be used to derive optimality properties for CIs analogous to those of tests. The general correspondence between a $(1 - \alpha)$ confidence intervals and tests is this: the confidence interval contains the values that would not be rejected by the given test at the specified level of significance (Neyman 1935); they would not be rejected because they would not be statistically significant (from the observed x_0) at significance level α, by the corresponding test. Consider test $T(\alpha)$. It follows that the $(1 - \alpha)$ one-sided interval corresponding to test $T(\alpha)$ is $\alpha > \bar{X} - c_\alpha(\sigma\sqrt{n})$. In particular, the 97.5% confidence interval estimator corresponding to test $T(\alpha)$ is:

$$\mu > \bar{X} - 1.96(\sigma\sqrt{n}).$$

To grasp the duality, one must think not of a fixed null hypothesis, e.g., $\mu = 0$, but rather of different values for μ_0 that might have been tested. In particular, were the test of null hypothesis,

H_0: $\mu < (\bar{x}c_\alpha(\sigma\sqrt{n})$, H_0 would have been rejected at level α. Similarly, the 95% CI for μ corresponding to the two-sided test, $T(.05)$ is:

$$(\bar{X} - 1.96(\sigma\sqrt{n}) \leq \mu < \bar{X} - 1.96(\sigma\sqrt{n})).$$

These dualities will figure importantly in wave III.

Fisher's Criticism of Confidential Intervals: Fiducial Intervals

Calling $(1 - \alpha)$ the "confidence level" of the estimation procedure was infelicitous. It encourages the supposition that $(1 - \alpha)$ is the degree of confidence to be assigned the particular interval *estimate* formed, once $\bar{X}$ is instantiated with $\bar{x}$. That would be fallacious. Once the estimate is formed, either the true parameter is or is not contained in it. One can say only that the particular estimate arose from a procedure which, with high probability, $(1 - \alpha)$,

would contain the true value of the parameter, whatever it is.

Fisher, in what is regarded as one of the most puzzling episodes in philosophy of statistics, seemed to advocate this fallacious instantiation for certain contexts. Fisher (1955) claimed N-P confidence interval methods are guilty of violating the principles of deductive logic by allowing

$$P(\bar{x} - c_\alpha(\sigma/\sqrt{n}) \leq \mu < \bar{x} - c_\alpha(\sigma/\sqrt{n})) = 1 - \alpha \quad (1)$$

and yet upon observing a particular $\bar{x}$, denying that the probability holds for the resulting CI estimate:

$$(\bar{x} - c_\alpha(\sqrt{n}) \leq \mu < \bar{x} - c_\alpha(\sqrt{n})) \quad (2)$$

Fisher claimed that, at least in certain special cases, it *was* possible to assign a probability or "fiducial distribution" to the interval statement about μ, while keeping within the sampling distribution perspective, a move that Savage (1962) described as "an attempt to make the Bayesian omelet without breaking the Bayesian eggs." Although the possibility of nonfallaciously instantiating into statement (1), to arrive at (2), without introducing a prior probability distribution, has tantalized researchers in philosophy of statistics, Fisher's fiducial argument is generally regarded as a lapse, even by Fisher's most ardent admirers (Hacking 1965; Seidenfeld 1979).

The Second Wave

The set of issues that swirled around the philosophy of statistics debates from the early 1960s through the late 1970s echoed the earlier debates but reflected as well changing problems in philosophy of science, statistics, and the statistical practices in the social sciences. Foundational debates of this period are noteworthy for the amount of direct interactions between philosophers of science and statistics; as is in evidence in the two significant collections of Godambe and Sprott (1971) and Harper and Hooker (1976).

As the most impressive mathematical developments of N-P theory occurred in a decision-theoretic framework, generalized further by Wald (1950)—the Neyman-Pearson-Wald (NPW) approach—it was the behavioristic-decision paradigm that bore the brunt of criticism from philosophy. Critics aimed at two central features of the "accept-reject" behavioristic conception of N-P tests: first, the justification of tests in terms of low (long-run) error rates alone, and second, the function of tests as routine, mechanical, or automatic accept-reject routines. While these features, taken strictly, give a caricature of tests—even as their founders intended and used them—they are at the heart of the philosophical criticisms of N-P testing. Not all critics call for tools that are more inferential and less decision-theoretic; some complained that N-P theory was at best a halfway house to a full-blown decision theory, with explicit loss functions, and prior probabilities that would be combined with measures of evidence (see Decision Theory). Because critics from both these camps hold a degree of confirmation stance, while error statisticians look to probability for objective measures of reliability of procedures, the disputants often talk past each other.

Error Probability Principle Versus Likelihood Principle

Hacking (1965) framed the main lines of criticism by philosophers in charging "Neyman-Pearson tests as suitable for before-trial betting, but not for after-trial evaluation" (99). Analogous charges are put in terms of distinctions between "initial precision" versus "final precision," and "before-data vs. after data" evaluation. According to such "post-data criticisms," N-P tools license inferences that while satisfactory from the pre-data viewpoint, seem unsatisfactory according to one of the post-data measures of (absolute or relative) evidential strength. The more general point may be put as follows:

- Data sets $\boldsymbol{x}$ and $\boldsymbol{y}$ may have exactly the same evidential relationship to hypothesis H, on a given degree of support measure, yet warrant different inferences according to significance test reasoning because x and y arose from tests with different *error probabilities*.

Such charges have weight, of course, only to the extent that one accepts the particular degree of support measure involved, the most common being based on the likelihood function of H, often written $L(H; \boldsymbol{x})$, where $L(H; \boldsymbol{x}) = P(\boldsymbol{x}; H)$. There is often confusion about likelihoods. Unlike the probability function, which assigns probabilities to the *different* possible values of the random variable of interest X, under some *fixed* value of the parameter(s) such as μ, the likelihood function gives the probability (or density) of a *given* observed value of the sample under the different values of the unknown parameter(s) such as μ.

Hacking (1965) championed an account of comparative support based on his "law of likelihood": Data $\boldsymbol{x}$ support hypotheses H_1 more than H_2 if the latter is *more likely* than the former, i.e.,

$P(x; H_1) > P(x; H_2)$. When there are many hypotheses, one takes the one that maximizes the likelihood. A problem is that there is always the rival hypothesis that things had to turn out the way they did. If such an alternative can always be constructed, then it will be possible to find H less well supported than some other hypothesis, even if H is true. Hacking (1965) rejected this likelihood approach on these grounds, but likelihoodist accounts are advocated by others and remain the focus of active interest (Birnbaum 1961; Royall 1997).

The likelihood function has an important role in all of the statistical accounts, but for those who endorse the likelihood principle, likelihoods suffice to convey "all that the data have to say." That is the gist of the *likelihood principle*—a pivot point around philosophy of statistics discussions:

> According to Bayes's theorem, $P(x|\mu)$...constitutes the entire evidence of the experiment, that is, it tells all that the experiment has to tell. More fully and more precisely, if y is the datum of some other experiment, and if it happens that $P(x|\mu)$ and $P(y|\mu)$ are proportional functions of μ (that is, constant multiples of each other), then each of the two data x and y have exactly the same thing to say about the values of μ .... (Savage 1962)

By contrast, the error probabilist must consider, in addition, the sampling distribution of the likelihoods (under hypotheses of interest). Thus, as Savage (1962) argued, significance levels and other error probabilities all violate the likelihood principle, leading to one of the most crucial philosophical controversies.

Debate Over the Relevance of the Stopping Rule

The conflict between significance levels and the LP is often illustrated by a variation on the two-sided test $T(2\alpha)$: a random sample from a normal distribution with mean μ and standard deviation 1, that is, $X_i \sim N(\mu, 1)$; with H_0: $\mu = 0$, and H_1: $\mu \neq 0$. However, instead of fixing the sample size n in advance, n is determined by a *stopping rule*:

$$\text{Keep sampling until } |\bar{x}| \geq 1.96/\sqrt{n}.$$

The probability that this rule will stop in a finite number of trials is 1, regardless of the true value of μ; it is a *proper* stopping rule. Whereas with n fixed in advance, such a test has a type 1 error probability of .05, with this stopping rule, the actual significance level differs from, and is greater than .05. Significance levels are sensitive to the stopping rule; and there is considerable literature on error probability adjustments for "optional stopping," that is, on *sequential tests* (e.g., Armitage 1961). By contrast, since likelihoods are unaffected by this stopping rule, the LP proponent denies there is an evidential difference between the two cases. For some, this was yet further grounds to embrace a Bayesian account:

> The likelihood principle emphasized in Bayesian statistics implies,...that the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proved or disproved. (Edwards, Lindman, and Savage 1963, 193)

For others it only underscored the point raised by Pearson and Neyman, that "knowledge of [the likelihood ratio] alone is not adequate to insure control of the error involved in rejecting a true hypothesis" (Pearson and Neyman 1930, 106). The literature here is vast; at best one can list sources (beyond those already mentioned) with fairly broad citations (Cox and Hinkley 1974; Mayo and Kruse 2001).

The key difference between the two perspectives is that the holder of the LP considers the likelihood of the *actual* outcome, that is, just $d(\boldsymbol{x})$, whereas the error statistician considers the likelihoods of values *other than the one observed* in order to assess the properties of the test procedure. The calculation of error probabilities, the sampling distribution, all depend on the relative frequency of outcomes other than the one observed, for example, outcomes as or more statistically significant—the "tail area." This remains a pivot point around which controversy in philosophy of statistics revolves. It is not a matter of one side being right and the other wrong, it is a matter of holding different aims, which in turn grow out of different philosophies of statistics.

The Significance Testing Controversy

Morrison and Henkel (1970) stands as a hallmark to the foundational issues wrestled with by social and behavioral scientists of this period. Where philosophers directed most of their criticisms to N-P tests, the focus here tended to center on simple Fisherian significance tests that had been widely adopted in psychology and other social sciences. Chastising social scientists for applying significance tests in slavish and unthinking ways, contributors call attention to a cluster of pitfalls and fallacies of testing. These fallacies are at the center of the philosophical controversies in this and later waves:

(i) *Large N Problem*: With large enough sample size, an α significant rejection of H_0 can be

very probable, even if the underlying discrepancy from μ_0 is substantively trivial. In fact, for any discrepancy from the null, however small, one can find a sample size such that there is a high probability (as high as one likes) that the test will yield a statistically significant result (for any p-value one wishes). Nevertheless, as Rosenthal and Gaito (1963) document, statistical significance at a given level is often (fallaciously) taken as more evidence against the null the larger the sample size (n). In fact, it is indicative of *less* of a discrepancy from the null than if it resulted from a smaller sample size. The "large n problem" is also the basis for the "Jeffrey-Good-Lindley" paradox brought out by Bayesians: even a highly statistically significant result can, as n is made sufficiently large, correspond to a high posterior probability accorded to a null hypothesis (see Bayesianism). Some suggest adjusting the significance level as a function of n, others, introducing some measure of the size of the discrepancy or "effect size" indicated. These issues return in the third wave.

(ii) *Fallacy of Non-Statistically Significant Results*: Test T(α) fails to reject the null, when the test statistic fails to reach the cutoff point for rejection, that is, $d(\boldsymbol{x}_0) \leq c_\alpha$. A classic fallacy is to construe such a "negative" result as evidence of the correctness of the null hypothesis. The problem is that merely surviving the statistical test is too easy, occurs too frequently, even when the null is false. One can always find a sufficiently small discrepancy δ from the null such that the test has low power to detect it. Thus, it would be fallacious to regard insignificant results as evidence that the discrepancy is less than δ, much less that there is no discrepancy at all. With publishers demanding at least a .05 significant result for publication, many of these studies remain tucked away, the so-called "file-drawer problem" (Meehl 1990).

The Power Analytic Movement of the 1960s

In their attempt to inculcate the calculation of power in psychology Cohen (1988) and others began, in the 1960s, the "power analytic" movement. The attention to power, of course, was a key feature of N-P tests, but apparently the prevalence of Fisherian tests in the social sciences, coupled, perhaps, with the difficulty in calculating power, resulted in power receiving short shrift.

Although this was less well advertised, the power analysts used power not only for planning but for interpreting nonsignificant results post-data: If a non-statistically significant result occurred with a test with low power to detect discrepancies of interest, the power analysts urged, then such a nonsignificant result should not be taken to rule out such departures from the null. In so doing, one is codifying a means to avoid the fallacy of taking "no evidence against" the null as "evidence for" the null.

It may be surprising to include Neyman, but one finds just such a post-data use of power in the occasional papers of Neyman in the 1950s. In one, Neyman addresses Carnapian confirmation: "In some sections of scientific literature the prevailing attitude is to consider that once a test, deemed to be reliable, fails to reject the hypothesis tested, then this means that the hypothesis is "confirmed". Calling this "a little rash" and "dangerous," he claims "a more cautious attitude would be to form one's intuitive opinion only after studying the power function of the test applied" (Neyman 1955, 41).

One is advised to consider: (i) how large a discrepancy from the null is considered "important" or non-trivial on substantive grounds (to be determined by the tester) $\delta_{\text{non-trivial}}$, and (ii) the power of detecting a $\delta_{\text{non-trivial}}$ with the test actually used, for example, Power($T(\alpha)$,$\delta_{\text{non-trivial}}$). If the power is low, "the fact that the test failed to detect the existence of δ "does not mean very much. In fact, [$\delta_{\text{nontrivial}}$] may exist and have gone undetected" Neyman (1957b, 16). So here in Neyman are the basic outlines of the post-data "power analytic" movement, admittedly, largely lost in the standard decision-behavior model of tests.

However, even the post-data use of power retains an unacceptable coarseness: power is always calculated relative to the cutoff point $c_\square$ for rejecting H_0. Consider test T($\alpha = .025$), $\sigma = 1$, $n = 25$, and suppose $\delta_{\text{non-trivial}} = .2$ is deemed "substantively important". To determine if "it is a little rash" to take a nonsignificant result, say $d(\boldsymbol{x}) = -.2$, as reasonable evidence that $\delta < \delta_{\text{nontrivial}}$ (i.e., an important discrepancy is absent), one is to calculate POW(T($\alpha = .025$), $\delta_{\text{nontrivial}}$) which is only .16! But why treat the particular non-significant result the same no matter how close it is to μ_0 (i.e., 0)? In fact $P(d(\boldsymbol{x}) > -.2; .2) \approx .93$. That is, were μ as large as .2, the test very probably would have detected

a more significant result. This suggests that rather than calculating

$$P(d(X) > c_\alpha; \mu = .2), \qquad (A)$$

one should calculate

$$(B)P(d(X) > d(X_0); \mu = .2). \qquad (B)$$

Even if (A) is low, (B) may be high. Whether Neyman and Pearson did or would have endorsed this modification of the pre-data error probabilities is an open question. The issue reappears in the "reforms" of the third wave.

The Third Wave: Relativism, Reforms, Reconciliations

Statistics in Meta-Methodology

In the 1980s and 1990s statistical inference began to figure in rational reconstructions of scientific episodes, in appraising methodological rules e.g., the value of novel evidence, the prediction versus accommodation debate (e.g., Howson and Urbach 1989; Glymour 1980; Mayo 1991) and in attempts to solve classic philosophical problems, such as Duhem's problem (Howson and Urbach 1989). The recognition that science in general, and statistical inference in particular, involves subjective judgments and values, the statistical method, most often appealed to here is largely one or another subjective Bayesian account. One can explain historical cases wherein anomalies are blamed on background rather than a hypothesis H, some argue, by showing how plausible prior beliefs could still permit H to have a reasonably high posterior degree of belief. Others charge that the very flexibility Bayes's theorem offers in reconstructing cases as rational is to sidestep the question at hand: Which hypothesis *ought* to be blamed for an anomaly? (Mayo 1997; Worrall 1993).

Bayesian Advances and Controversy

The heat of the old debates is less in evidence in the third wave. For the most part statisticians are comfortable with an eclecticism, wherein different methods may be suitable for different functions, for example "pure" (Fisherian) tests in some cases, N-P "decision procedures" in others, along with good-sense, informal recommendations for their interpretation. To others, particularly nonstatistician practitioners (e.g., in psychology, ecology, medicine), the situation seems less one of joyful eclecticism, and more one of "unholy hybrids" yielding a mixture of ideas from N-P methods, Fisherian tests, and Bayesian accounts that is "inconsistent from both perspectives and burdened with conceptual confusion" (Gigerenzer 1993, 323). Because increasingly philosophers of science come to these issues by way of subject matter fields, they are more likely to be users of the latest methods rather than occupy their historical role as outside critic.

The use of Bayesian methods has grown exponentially both because of the philosophical problems with error statistical methods as well as the development of effective computational tools such as a Markov Chain Monte Carlo (MCMC). The rise in statistical computer packages means that Bayesian and non-Bayesian methods are readily available, encouraging the practitioner to view them as simply enriching the statistical toolkit rather than as reflecting different perspectives on philosophical foundations. In this sense the use of high-powered statistical tools increases the distance between the use and philosophical foundations of the methods. But when competing interpretations arise, as they often do, the philosophical questions from the first and second waves re-emerge. Most especially are debates about the role and justification of Bayesian prior probabilities. Operating with mathematically convenient priors is common but, as Bayesians are well aware, more is needed to justify them. One important argument put forward shows that with sufficient data, posterior probabilities will converge even if they are based on different priors (see Bayesianism). As Kyburg (1993, 146) shows, however, for any body of data there are non-extreme prior probabilities that will result in posteriors that differ by as much as one wants.

A related argument defending the use of priors shows that it is possible to ascertain the influence different priors may have, and so long as the posterior remains relatively insensitive the Bayesian inference is *robust* to the prior. A question that arises is this: if when the choice of prior is found to matter one must seek a different procedure, and if there are sufficient data such that the choice of prior scarcely matters, then why is the prior relevant at all? Does not this revert to the goal that drove Neyman, namely, to find procedures whose validity does not depend on the priors? The appeal of error statistical methods, despite problems, is that they apply for the kinds of uncertain cases scientists often face. Granted it is appealing to enlist the beliefs of "experts," but the question is how to retain the ability to critique and hold them accountable—a growing concern in evidence-based policy. Error probabilities can be calibrated against empirical frequencies, but can one equally well calibrate the opinion of the experts?

Reforms Within Error Statistics

There is an extensive movement to retain error statistical tools and yet reform them in order to avoid the well-known fallacies and shortcomings. The significance test fails to convey the effect of discrepancy warranted, and thus many journals require they be supplemented with measures of effect size. The most fruitful idea seems to be to appeal to two sided CI estimation procedures, even in interpreting the one-sided test $T(\alpha)$.

Consider interpreting non-significant results. Since all elements of the CI "fit" or are consistent with the outcome at the given level, the interpreter is deterred from thinking there is evidence for 0. But, as critics note, this will not go far enough to block fallacies of acceptance in general. For example, the $(1 - 2\alpha)$ CI for the parameter μ in test $T(\alpha)$ with $\alpha = .025$ is: $[\bar{x} - 1.96(\sigma/\sqrt{n}), \bar{x}1.96(\sigma/\sqrt{n})]$ $\sigma = 1, n = 25, (\sigma/\sqrt{n}) = .2$. Outcome $\bar{x} = .39$ just fails to reject H_0 at the .025 level, and correspondingly 0 is included in the two-sided 95% interval: $(-.002 < \mu < .782)$ (see duality between tests and CIs, above). Consider now the inference $\mu < \mu_1$ for μ_1 within the CI, say, $\mu < 0.2$. The hypothesis $\mu < 0.2$ is non-rejectable by the test—it is a survivor, as it were. But the construal is dichotomous: In or out, plausible or not; all values within the interval are *on par*, as it were (Mayo 1996). This does not adequately prevent fallacious interpretations of non-significance (fallacies of acceptance). Although $\bar{x}$ is not sufficiently greater (or less) than any of the μ values in the confidence interval to reject them at the α-level, this does not imply there is evidence *for each* of the values in the interval (Mayo and Spanos 2005).

Severity Assessments

The power analyst would seem to do better here. For each value of μ_1 in the confidence interval, there would be a different answer to the question: What is the power of the test against μ_1? Thus the power analyst makes distinctions that the CI interval theorist does not. The power analyst blocks the inference $\mu < 0.2$ since POW(T($\alpha = .025$), .2) is low (.16). But, as seen in the second wave, there is an important weakness of the use of power to avoid fallacies of acceptance. Were the result not $\bar{x} = .39$, but rather $x = -.2$, the test again fails to reject H_0, but the power analyst, looking just at $c_\alpha = 1.96$ is led to the same assessment denying there is evidence for $\mu < 0.2$. (power analysts commonly recommend a power of .8 as high). Although the "prespecified" power is low, .16, it seems clear that the interpretation, post-data, should reflect the actual outcome, and there is a high probability for a more significant result than the one attained, were μ as great as 0.2! Rather than construe "a miss as good as a mile," parity of logic suggests that the post-data power assessment should replace the usual calculation of power against μ_1:

$$\text{POW}(\text{T}(\alpha), \mu_1) = P(d(\mathbf{X}) > c_\alpha; \mu = \mu_1),$$

with what might be called the *power actually attained* or, to have a distinct term, the *severity (SEV)*:

$$\text{SEV}(\text{T}(\alpha), \mu_1) = \text{P}(d(\mathbf{X}) > d(x_0); \mu = \mu_1),$$

where $d(x_0)$ is the observed (nonstatistically significant) result (Mayo and Cox 2005). SEV(T(α), $d(x_0)$, $\mu < \mu_1$) is a shorthand for "the severity of the test which $\mu < \mu_1$ has passed on the basis of the insignificant result $d(x_0)$ from test T(α)." This is the post-data measure of a test's *severity* for detecting discrepancies as large as $\gamma = \mu_1 - \mu_0$. Since T(α)'s probativeness would be even higher for greater values of μ, it follows that SEV(T(α), $\mu < \mu_1$) > P($d(\mathbf{X}) > d(x_0)$; $\mu = \mu_1$) Mayo and Spanos 2005).

The philosophical position here is that error probabilities serve a function in a post-data interpretation of statistical inferences, by characterizing the probativeness of the particular test result with respect to a particular interpretation or particular inference on may wish to consider: Pre-data, one is balancing the two types of errors; but post-data, the concern shifts to evaluating if particular inferences are warranted. Figure 1 compares power and severity.

Conversely, for any non-significant result from test T(α), one may find the value of μ against which the test has high severity, say .975. This is solved by $\mu_1 = \bar{x} + 1.96\sigma_{\bar{x}}$, which is noticed to be the same value as the upper bound of a two-sided .95 level CI, μ. However, unlike the use of CIs, the severity analysis discriminate between inferences $\mu < \mu_1$ for different values of μ_1 within the interval. The computations related to delineating a series of observed CIs at different levels can be found in Kempthorne's "consonance intervals" (Kempthorne and Folks, 1971) and "confidence curves," "*p*-value functions" (Birnbaum 1961; Poole 1987). These strategies are motivated by the desire to move away from (i) having to choose a particular confidence level (or corresponding *p*-value), (ii) the dichotomous, "up"/"down" interpretations of tests. There would appear to be an important difference with these approaches, at least in emphasis. If one is thinking of values "consistent" with the observed data x_0, then a value μ' near the center of the CI is more in accord with x_0 than is μ'' near the upper CI bound;

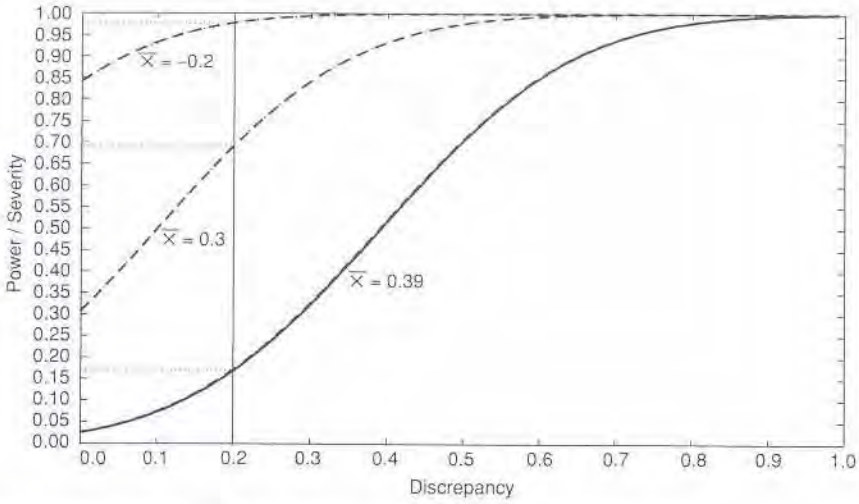

Fig. 1. The graph shows that whereas POW($\bar{T}(.025), \mu_1 = .2) = .168$, irrespective of the value of $d(x_0)$ (or $\bar{x}$); see solid curve, the severity evaluations are data-specific: for $d(x_0) = 1.95$(or $\bar{x} = .39$), SEV($\bar{T}(.025), \mu < .2) = .171$; for $d(x_0) = 1.50$(or $\bar{x} = .30$), SEV($\bar{T}(.025), \mu < .2) = .691$, and for $d(x_0) = -1.0$(or $\bar{x} = -.2$), SEV($\bar{T}(.025), \mu < .2) = .977$.

however the inference $\mu < \mu''$ has passed a more probative test than has $\mu < \mu'$.

Fallacies of Rejection: The Large* n *Problem

While with a nonsignific*ant* result, the concern is erroneously inferring that a discrepancy from μ_0 is absent; with a *significant* result x_0, the concern is erroneously inferring that it is present. Rejection need not be discussed separately here (see Mayo 1996), since for any H: Sev($\neg H$) = 1 − Sev(H), it follows for the particular case of H_1: $\mu > \mu_1$ Sev ($\mu > \mu_1$) = 1 − Sev ($\mu < \mu_1$) = 1− (actual power at ($\mu = \mu_1$).

The "large n" problem already made its splash in the second wave: With large enough sample size, an α significant rejection of H_0 can be very probable for any discrepancy α from μ_0, even if it is *substantively* trivial. Utilizing the severity assessment, an α-significant difference with n_1 passes $\mu > \mu_1$ less severely than with n_2 where $n_1 > n_2$.

Figure 2 compares test T(α) with three different sample sizes: $n = 25$, $n = 100$, $n = 400$, denoted by T(α, n); where in each case $d(x_0) = 1.96$ – reject at the cutoff point.

More generally, if two (otherwise identical) tests with different sample sizes give rise to rejections of H_0 at the same p-value, the result from the smaller sample experiment indicates a greater extent of a discrepancy from H_0 than from the larger. This immediately scotches the "large n problem," and simultaneously provides a way to supply p-values with assessments of population discrepancy (or effect size) that can be compared across different tests.

P-values and Bayesian Posteriors

Severity is an error probability calculation *based on* the actual data (and inference of interest) but it must be distinguished from what has sometimes been called the *conditional* "error probability" understood as a posterior probability. The most well-known fallacy in interpreting significance tests is to equate the p-value with a posterior probability on the null hypothesis. The p-value assessment refers only to the sampling distribution of the test statistic $d(X)$; and there is no use of priors. The Jeffrey-Good-Lindley "paradoxical" examples (see above) shows that attaining a fixed p-value, with a sufficiently large n, can correspond to large posterior probabilities for H_0. More recent work generalizes the result (Berger and Sellke 1987). Although from the degree-of-confirmation perspective, it follows that p-values come up short as a measure of evidence, the significance testers balk at the fact that use of the recommended priors can result in highly significant results being construed as no evidence against the null—or even evidence for it! An interesting twist in recent work is to try to "reconcile" the p-value and the posterior (e.g., Berger 2003).

The conflict between p-values and Bayesian posteriors often considers the familiar example of the two-sided T(2α)test, H_0: $\mu = \mu_0$ versus H_1:$\mu \neq \mu_0$. The difference between p-values and posteriors

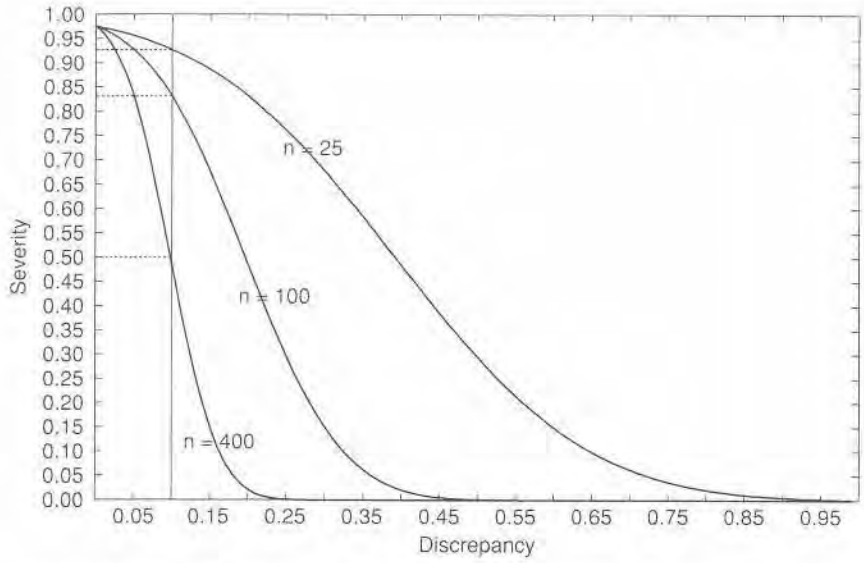

Fig. 2. In test $T(\alpha)$, $(H_0 : \mu \leq 0$ against $H_1 : \mu > 0$, and $\sigma = 1)$, $\alpha = .025$, $c_\alpha = 1.96$ and $d(x_0) = 1.96$. Inference under evaluation:
$\mu > 0.1$: $SEV(T(\alpha, 25), \mu = 0.1) = .93$; $SEV(T(\alpha, 100), \mu = 0.1) = .83$; $SEV(T(\alpha, 400), \mu = 0.1) = .5$

are far less marked with one-sided tests (e.g., Pratt 1977). "If $n = 50$ one can classically 'reject H_0 at significance level $p = .05$,' although $P(H_0|x) = .52$ (which would actually indicate that the evidence favors H_0)" (Berger and Sellke 1987, 113, replace Pr with P for consistency). Thus, data that the significance tester would regard as evidence against H_0, would, on the Bayesian construal being advocated actually indicate that the evidence favors H_0. If $n = 1000$, a result statistically significant at the .05 level leads to a posterior to the null of .82!

What makes the example so compelling to many is its use of an "impartial" or "uninformative" Bayesian prior probability assignment of .5 to H_0, the remaining .5 probability being spread out over the alternative parameter space, e.g., as recommended by Jeffreys (1939). Others charge that the problem is not p-values but the high prior. Moreover, the "spiked concentration of belief in the null" is at odds with the prevailing view "we know all nulls are false." Note too the conflict with CI reasoning since *0* is outside the corresponding CI.

Some examples strive to keep within the frequentist camp: to construe a hypothesis as a random variable, it is imagined that there is random sampling from a population of hypotheses, some proportion of which are assumed to be true. The percentage "initially true" serves as the prior probability for H_0. This gambit is common across all philosophy of statistics literature, and yet it commits a fallacious instantiation of probabilities:

> 50% of the null hypotheses in a given pool of nulls are true. This particular null hypothesis H_0 was randomly selected from this pool. Therefore $P(H_0$ is true$) = .5$.

Faced with conflicts between error probabilities and Bayesian posterior probabilities, the error probabilist would conclude that the flaw lies with the latter measure. This is precisely what Fisher argued, and it seems fitting to end up this retrospective with a return to him.

Discussing a test of the hypothesis that the stars are distributed at random, Fisher takes the low p-value (about 1 in 33,000) to "exclude at a high level of significance any theory involving a random distribution" (Fisher 1956, 42). Even if one were to imagine that H_0 had an extremely high prior probability, Fisher continues—never minding "what such a statement of probability a priori could possibly mean"—the resulting high posteriori probability to H_0, he thinks, would only show that "reluctance to accept a hypothesis strongly contradicted by a test of significance" (ibid, 44) "is not capable of finding expression in any calculation of probability a posteriori" (ibid, 43). It is important too to recognize that sampling theorists do not deny there is ever a legitimate frequentist prior probability distribution for a statistical hypothesis: one may consider hypotheses about such distributions and subject them to probative tests. Indeed, if one were to consider the claim about the a priori probability to be itself a

hypothesis, Fisher suggests, it would be rejected by the data!

Concluding Comment

Underlying the central points of controversy in the three waves of philosophy of statistics lie two contrasting philosophies of the role of probability in statistical inference. In one tradition, probability is used to provide a post-data assignment of degree of probability, confirmation, support or belief in a hypothesis (e.g., Bayesian and likelihood accounts); while in a second, probability is used to assess the probativeness, reliability, trustworthiness, or severity of a test or inference procedure (e.g., significance tests, N-P tests, CI). This basic contrast in underlying aims corresponds to conflicting principles for appraising methods: satisfying the likelihood principle, as opposed to controlling error probabilities. Whether statistical methodology should be regarded as supplying different tools depending on the task at hand, or whether the different methods can or should be reconciled in some way, are likely to remain questions of debate for a good while longer.

DEBORAH G. MAYO

The author acknowledges the helpful input of Aris Spanos and D. R. Cox.

References

Armitage, P. (1961), Contribution to the discussion in Smith, C.A.B., "Consistency in Statistical Inference and Decision." *Journal of the Royal Statistical Society, B* 23: 1–37

Berger, J. (2003), "Could Fisher, Jeffreys and Neyman Have Agreed?" *Statistical Science* 18: 1–12.

Berger, J. O., and T. Sellke (1987), "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence." *Journal of the American Statistical Association* 82: 112–122.

Birnbaum, A. (1961), "Confidence Curves: An Omnibus Technique for Estimation and Testing Statistical Hypotheses." *Journal of the American Statistical Association* 56: 246–249.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cox, D. R., and D. V. Hinkley (1974), *Theoretical Statistics*. London: Chapman and Hall.

Edwards, W., H. Lindman, and L. Savage (1963), "Bayesian Statistical Inference for Psychological Research." *Psychological Review* 70: 193–242.

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

——— (1955), "Statistical Methods and Scientific Induction." *Journal of the Royal Statistical Society B* 17: 69–78.

——— (1956), *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

Friedman, D. (1995), "Some Issues in the Foundation of Statistics." *Foundations of Science* 1: 19–39.

Gigerenzer, G. (1993), "The Superego, the Ego, and the Id in Statistical Reasoning," in G. Keren, and C. Lewis (eds.), *A Handbook of Data Analysis in the Behavioral Sciences: Methodological Issues*. Hillsdale, NJ: Erlbaum, 311–339.

Glymour, C. (1980), *Theory and Evidence*. Princeton, NJ: Princeton University Press.

Godambe, V., and D. Sprott (eds.) (1971), *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston.

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

Harper, W., and C. Hooker (eds.) (1976), *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science* (Vol. 2). Dordrecht, Netherlands: D. Reidel.

Howson, C., and Urbach, P. (1989), *Scientific Reasoning: The Bayesian Approach*. LaSalle, IL: Open Court.

Jeffreys, H. (1939), *The Theory of Probability*. Oxford: Oxford University Press.

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

Kyburg (1993), "The Scope of Bayesian Reasoning," in Hull, D. Forbes, L. and Okruhlik, K. (eds.), *PSA*, vol. 2. East Lansing, MI: PSA, 139–152.

Lehmann, E. L. (1993), "The Fisher and Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?" *Journal of the American Statistical Association* 88: 1242–1249.

Mayo, D. G. (1983), "An Objective Theory of Statistical Testing." *Synthese* 57: 297–340.

——— (1991), "Novel Evidence and Severe Tests." *Philosophy of Science* 58: 523–552.

——— (1996), *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

——— (1997), "Duhem's Problem, The Bayesian Way, and Error Statistics, or 'What's Belief Got To Do With It'?" and "Response to Howson and Laudan." *Philosophy of Science* 64: 222–224 and 323–333.

Mayo, D. G., and D. R. Cox (2005), "Frequentist Statistics as a Theory of Inductive Inference." Proceedings of the Second Erich L. Lehmann Symposium. *Institute for Mathematical Statistics Lecture Notes—Monograph Series* 70, forthcoming.

Mayo, D. G., and M. Kruse (2001) "Principles of Inference and Their Consequences," in D. Cornfield and J. Williamson (eds.), *Foundations of Bayesianism*. Dordrecht, Netherlands: Kluwer Academic Publishers, 381–403.

Mayo, D. G., and A. Spanos (2005), "Severe Testing as a Basic Concept in the Neyman-Pearson Philosophy of Induction." *British Journal of the Philosophy of Science* forthcoming.

Meehl, P. E. (1990), "Why Summaries of Research on Psychological Theories are Often Uninterpretable." *Psychological Reports* 66: 195–244.

Morrison, D., and R. Henkel (eds.) (1970), *The Significance Test Controversy*. Chicago: Aldine.

Neyman, J. (1935), "On the Problem of Confidence Intervals." *Annals of Mathematical Statistics* 6: 111–116.

——— (1956), "Note on an Article by Sir Ronald Fisher." *Journal of the Royal Statistical Society, B (Methodological)* 18: 288–294.

——— (1955), "The Problem of Inductive Inference." *Communications on Pure and Applied Mathematics* 8: 13–45.

——— (1957a), "Inductive Behavior as a Basic Concept of Philosophy of Science," *Revue de l'Institut International de Statistique* 25, 7–22.
——— (1957b), "The Use of the Concept of Power in Agricultural Experimentation," *Journal of the Indian Society of Agricultural Statistics* IX: 9–17.
Neyman, J., and E. S. Pearson (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society, A* 231: 289–337.
Pearson, E. S. (1950), "On Questions Raised by the Combination of Tests Based on Discontinuous Distributions," *Biometrika* 37: 383–398.
——— (1955), "Statistical Concepts in Their Relation to Reality," *Journal of the Royal Statistical Society B* 17: 204–207.
Pearson, E. S., and J. Neyman (1930), "On the Problem of Two Samples," *Bulletin of the Academy of Political Science* 73–96, as reprinted in J. Neyman and E. S. Pearson (1967), 99–115.
Poole, C. (1987), "Beyond the Confidence Interval," *American Journal of Public Health* 77: 195–199.
Rosenthal, R., and J. Gaito (1963), "The Interpretation of Levels of Significance by Psychological Researchers," *Journal of Psychology* 64: 725–739.
Royall, R. (1997), *Statistical Evidence: A Likelihood Paradigm*. London: Chapman and Hall.
Salmon, W. (1967), *The Foundations of Scientific Inference*. Pittsburgh: University of Pittsburgh Press.
Savage, L. (ed.) (1962), *The Foundations of Statistical Inference: A Discussion*. London: Methuen.
Wald, A. (1950), *Statistical Decision Fuctions*. New York: Wiley.
Worrall, J. (1993), "Falsification, Rationality, and the Duhem Problem," in J. Earman, A. Janis, G. Massey, and N. Rescher (eds.), *Philosophical Problems of the Internal and External Worlds, Essays on the Philosophy of Adolf Gruenbaum*. Pittsburgh: University Pittsburgh Press, 329–370.

See also **Bayesianism; Confirmation Theory; Decision Theory; Probability**

STATISTICAL MECHANICS

See **Kinetic Theory**

STRONG PROGRAM

See **Social Constructionism**

SUPERVENIENCE

The term 'supervenience,' as appropriated by the philosophical community, denotes a relation between two families of properties. Roughly stated, the *A*-properties supervene on the *B*-properties just in case there can be no difference in *A*-properties without some difference in *B*-properties. Equivalently, if two things are exactly alike in *B*-properties, they must be exactly alike in *A*-properties.

A simple and uncontroversial example of supervenience may help fix ideas: the case of aesthetic and nonaesthetic properties. If any two objects are exactly alike with regard to their nonaesthetic properties, they must be exactly alike with regard to their aesthetic properties; indiscernibility in nonaesthetic properties requires indiscernibility in aesthetic properties. Other normative properties