# STATISTICAL SIGNIFICANCE AND THE BURDEN OF PERSUASION

DAVID H. KAYE*

## I

## INTRODUCTION

In most endeavors concerned with the acquisition of knowledge, quantitative information is welcomed. In law, however, it appears sometimes that scientific or numerical evidence makes cases harder, not easier. Nevertheless, there are many cases and administrative proceedings, in such areas as environmental law, food and drug regulation, and civil rights, in which statistical data obtained by observation or experiment are readily accepted as assisting in the proper resolution of disputed issues of fact.[1] When courts or administrators confront scientific and statistical evidence in these proceedings, they are not always certain of how to weigh the evidence or whether they should, or must, rely on the standards for proof that scientists apply in evaluating statistical hypotheses.

Two decisions of the Supreme Court illustrate this uncertainty. In *Castaneda v. Partida,*[2] a grand jury discrimination case, the Court, acting as its own statistician, computed a statistic known as the "standard deviation."[3] The Court found this computation highly probative of discrimination in light of the "general rule" that "if the difference between the expected value and the observed number is greater than two or three standard deviations, the hypothesis that the jury drawing was random would be suspect to a social scientist."[4] One would not have thought that this reference, in a single footnote, to a standard of proof popular among social scientists would be read as commanding that the same standard be determinative when evaluating statistical evidence in court. Yet, in *Hazelwood School District v. United States,*[5] an employment discrimination case decided the same year, the Court dropped the qualifying language about social science and noted that, under the "precise" methodology delineated in *Castaneda,* a disparity of slightly less than two standard deviations was not "suspect."[6] In the wake of these opinions,

* Professor of Law, Arizona State University. The author wishes to thank Hans Zeisel for commenting on a draft of this article.

[1] *See generally* C. MCCORMICK, A HANDBOOK ON THE LAW OF EVIDENCE §§ 208-211 (3d ed. E. Cleary 1984).

[2] 430 U.S. 482 (1977).

[3] *Id.* at 496 n.17. The standard deviation measures the variability or dispersion of a set of numbers. If all of the numbers are the same, the standard deviation is zero. If most of the numbers are far from their mean, the standard deviation is large.

[4] *Id.*

[5] 433 U.S. 299 (1977).

[6] *Id.* at 309 n.14. For criticism of the *Hazelwood* Court's manipulation of the statistics, see Kaye, *Statistical Evidence of Discrimination,* 77 J. AM. STATISTICAL A. 773 (1982); Smith & Abram, *Quantitative*

litigants in discrimination cases undertake intensive searches for numbers that can be translated into standard deviations, and plaintiffs have come to treasure discrepancies that amount to at least two, and preferably three, of these standard deviations.[7]

This quest for "statistical significance" is not confined to discrimination cases. In any case involving statistical proof, the proponent of the evidence understandably covets testimony that the data is "significant." Although there are grounds to question whether such testimony as to "significance" should even be admissible,[8] this article confines attention to one issue—the relationship between the putative scientific standard of proof and the legal standards of proof. Section II describes these legal standards, first in the conventional language of the law and then in the terminology of statistical decision theory. Section III describes the technical concept of statistical significance and shows the impossibility, in general, of equating statistical significance with legally satisfactory proof. What follows from this analysis is, I hope, the beginning of a clearer understanding of how statistical methods for measuring the probative force of data can help the trier of fact decide whether the proponent of the data has fulfilled the appropriate burden of persuasion.[9]

## II

## THE LEGAL STANDARDS OF PROOF

### A. The Legal Formulas

Witnesses testify. Lawyers argue. Courts decide. With these rituals, many things are accomplished.[10] In part, all the participants contribute in an effort to reconstruct the past. Yet, the success of such efforts rarely can be known. In any seriously disputed case, some uncertainty remains. Even so, a verdict must be returned, and this verdict may have immediate and dramatic consequences for the ebb and flow of money and for the activities, if not the very lives, of persons or organizations.

*Analysis and Proof of Employment Discrimination*, 1981 U. ILL. L.F. 33, 52-53; Kaye, Book Review, 80 MICH. L. REV. 833, 838-41 (1982) (reviewing D. BALDUS & J. COLE, STATISTICAL PROOF OF DISCRIMINATION (1980)).

[7] In the employment discrimination field, cases involving expert testimony structured to replicate the "standard deviation analysis" of *Castaneda* and *Hazelwood* are becoming legion, so much so that the failure to pursue this analysis leads some courts to discredit the statistical evidence. *See, e.g.,* Hill v. K-Mart Corp., 699 F.2d 776, 780 n.7 (5th Cir. 1983). All too often, conclusory statements as to whether the number of standard deviations exceed two or three are given greater weight than the number itself. *See, e.g., id.*; EEOC v. H.S. Camp & Sons, Inc. 29 Empl. Prac. Dec. ¶ 32,930, at 26,369 (M.D. Fla. 1982) ("an average statistical disparity of 1.39 standard deviations" held to be "far below the standard deviation level that has been established as constituting a gross statistical disparity," and an "average statistical disparity of only 2.87 standard deviations" dismissed as falling below this level). For additional citations, see D. BALDUS & J. COLE, STATISTICAL PROOF OF DISCRIMINATION (1980); W. CONNOLLY & D. PETERSON, USE OF STATISTICS IN EQUAL EMPLOYMENT OPPORTUNITY LITIGATION (1980).

[8] I am preparing a more comprehensive analysis of hypothesis and significance testing that develops such an argument.

[9] The forthcoming paper mentioned in note 8 offers some suggestions as to how the procedures used in hypothesis testing might be adapted to improve legal factfinding.

[10] *See, e.g.,* Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process,* 84 HARV. L. REV. 1329 (1971).

To facilitate and structure these decisions, and to cope with the factual uncertainty that remains after witnesses testify and lawyers argue, the law of evidence imposes on one or another party a component of the burden of proof known as the "burden of persuasion."[11] For example, in a murder case the prosecution must prove "beyond a reasonable doubt" that the defendant did in fact take the life of another human being. In different types of cases, a different showing may be required to "carry" the burden of proof.[12] Thus, in most civil litigation a mere "preponderance of the evidence" suffices. In some civil cases, however, a more stringent standard will be applied. In these "quasi-criminal" matters, one must adduce "clear and convincing" or "clear, convincing, and unequivocal" proof.[13]

The courts recognize that these varying burdens of persuasion reflect the

[11] The burden of proof usually is said to involve two components — the burden of production and the burden of persuasion. *See* J. THAYER, A PRELIMINARY TREATISE ON EVIDENCE AT THE COMMON LAW 355-59 (1898); James, *Burdens of Proof,* 47 U. VA. L. REV. 51 (1961). Plaintiffs and prosecutors generally have the burden of proving all the factual elements essential to their case under the applicable substantive law. In a prosecution for murder, for instance, the state must prove (among other things) that the defendant took the life of another human being. In a product liability action involving, for example, a runaway tractor-type lawnmower that allegedly chewed its way into a neighbor's house, the plaintiff must prove (among other things) that the exuberant mower was defective. This burden of proof compels the party to come forward with evidence sufficient to convince the judge that a reasonable jury *could* conclude, on the strength of this evidence, that the requisite facts have been established. This is the burden of production, also known as the "burden of evidence," 31A C.J.S. *Evidence* § 103 (1964), and the "duty of going forward." J. THAYER, *supra,* at 355. If the party with the burden fails to establish a prima facie case because the evidence is inadequate to permit a reasonable person to believe that the requisite facts exist, then the judge should direct a verdict and not permit the burden of production to shift to the defendant. *See, e.g.,* Texas Dep't of Community Affairs v. Burdine, 450 U.S. 248, 254 (1981) (plaintiff's prima facie case of disparate treatment shifts the burden of production to the defendant and "[i]f the trier of fact believes the plaintiff's evidence, and if the employer is silent . . . the court must enter judgment for the plaintiff . . . ."). Placing the burden of proof on one party means that this party must do more than introduce a bare minimum of evidence. He must convince the jury of the fact in dispute. The burden of persuasion specifies how convincing his evidence must be. *See* McNaughton, *Burden of Production of Evidence: A Function of the Burden of Persuasion,* 68 HARV. L. REV. 1382 (1955). Wigmore used the phrase the "risk of nonpersuasion." 9 J. WIGMORE, EVIDENCE IN TRIALS AT COMMON LAW § 2485 (1940); *see also* Winter, *The Jury and the Risk of Nonpersuasion,* 5 LAW & SOC'Y REV. 335 (1971).

[12] *But see* Underwood, *The Thumb on the Scales of Justice: Burdens of Persuasion in Criminal Cases,* 86 YALE L.J. 1299, 1300 n.4 (1977):

> The metaphor of carrying burdens is confusing for two reasons. First, it suggests, inaccurately, that the law is concerned with the extent of a party's labors, when in fact the evidence that satisfies the burden may be introduced by either party. Second, it seems implausible to tell the factfinder to reach a decision, attach a probability estimate to that decision, and then use a legal rule about burdens to translate the decision into a verdict, sometimes translating a decision for *A* into a verdict for *B*. Once he has decided, it seems odd to use a rule to tell him to change his decision. It is more sensible, and truer to the purpose of the rule, to tell him not to decide when he cannot be sure enough, and then to provide a legal rule to make the decision for him in such cases.

This article takes a somewhat different view of what is "sensible." Rather than distinguishing between "decisions" and "verdicts" and treating the burden of persuasion as a direction to the factfinder to "regard the dispute as too close for decision," Underwood, *supra,* at 1300, this article interprets the rule as telling factfinders how they should translate probability estimates into decisions. I do not imply, however, that jurors actually decide cases by making subjective probability statements, then converting them into verdicts. Statistical decision theory and signal detection theory (of which Professor Underwood, *supra,* at 1331 n.93, seems to approve) are introduced in Section B solely to provide a heuristic or normative model of what the burden of persuasion should mean if it is to further certain goals. *See* Kaye, *Probability Theory Meets Res Ipsa Loquitur,* 77 MICH. L. REV. 1456 (1979); Lempert, *Modeling Relevance,* 75 MICH. L. REV. 1021 (1977).

[13] *See* C. MCCORMICK, HANDBOOK ON THE LAW OF EVIDENCE §§ 340-341 (E. Cleary 2d ed. 1972). The phrase "quasi-criminal" appears in Addington v. Texas, 441 U.S. 418, 424 (1979).

relative costs of erroneous verdicts. In *Speiser v. Randall,*[14] the Supreme Court held that a state property tax exemption conditioned on a loyalty oath unconstitutionally imposed a burden on claimants to prove their loyalty. In dictum, Justice Brennan, writing for the majority, explained that:

> "[t]here is always in litigation a margin of error, representing error in factfinding which both parties must take into account. Where one party has at stake an interest of transcending value — as a criminal defendant his liberty — this margin of error is reduced as to him by the process of placing on the other party the burden of . . . persuading the factfinder at the conclusion of the trial of his guilt beyond a reasonable doubt.[15]

In other words, the "interest of transcending value"—the liberty of the person—makes the cost of an erroneous verdict for the state greater than the cost of an erroneous verdict for the defendant and underlies the requirement of proof beyond a reasonable doubt in criminal cases.

In addressing the constitutional foundations of the beyond a reasonable doubt requirement as applied to a juvenile delinquency proceeding, Justice Harlan elaborated on this theme in his concurring opinion in *In re Winship.*[16] Harlan wrote:

> In a lawsuit between two parties, a factual error can make a difference in one of two ways. First, it can result in a judgment in favor of the plaintiff when the true facts warrant a judgment for the defendant. The analogue in a criminal case would be the conviction of an innocent man. On the other hand, an erroneous factual determination can result in a judgment for the defendant when the true facts justify a judgment in plaintiff's favor. The criminal analogue would be the acquittal of a guilty man.[17]

He explicitly reasoned that the appropriate burden of persuasion should turn on the relative seriousness of the possible errors:

> The standard of proof influences the relative frequency of these two types of erroneous outcomes. If, for example, the standard of proof for a criminal trial were a preponderance of the evidence rather than proof beyond a reasonable doubt, there would be a smaller risk of factual errors that result in freeing guilty persons, but a far greater risk of factual errors that result in convicting the innocent. Because the standard of proof affects the comparative frequency of these two types of erroneous outcomes, the choice of the standard to be applied in a particular kind of litigation should, in a rational world, reflect an assessment of the comparative social disutility of each.[18]

The Justice illustrated the relationship between the "comparative social disutility" and the burden of persuasion by considering the preponderance of the evidence standard:

> In a civil suit between two private parties for money damages, for example, we view it as no more serous in general for there to be an erroneous verdict in the defendant's favor than for there to be an erroneous verdict in the plaintiff's favor. A preponderance of the evidence standard therefore seems peculiarly appropriate for, as explained most sensibly, it simply

[14] 357 U.S. 513 (1958).

[15] *Id.* at 525-26 (dictum).

[16] 397 U.S. 358 (1970). *Winship* held the reasonable doubt standard constitutionally mandated in an adjudication of delinquency based on conduct that would constitute a crime for an adult. *Id.* at 368. Only Mr. Justice Black questioned the constitutional basis for the requirement of proof beyond a reasonable doubt in a criminal trial of an adult. *Id.* at 377 (Black, J., dissenting). The majority was of the opinion that the "Due Process Clause protects the accused against conviction except upon proof beyond a reasonable doubt of every fact necessary to constitute the crime with which he is charged." *Id.* at 364.

[17] *Id.* at 370-71 (Harlan, J., concurring).

[18] *Id.* at 371.

> requires the trier of fact 'to believe that the existence of a fact is more probable than its nonexistence . . . .[19]

The Court has had occasion to advert to Justice Harlan's "comparative social disutility" analysis in several recent cases. For example, in holding that due process demands at least proof by clear and convincing evidence, rather than a preponderance of the evidence, for an involuntary civil commitment, a unanimous Court observed in *Addington v. Texas*[20] that the preponderance of the evidence standard requires the litigants to "share the risk of error in roughly equal fasion."[21] Last year, in *Santosky v. Kramer,*[22] the Court held a preponderance of the evidence standard violative of due process when applied to terminate parental rights on the basis of a finding of permanent neglect. Although the Court divided sharply over the result in the case,[23] all the Justices seemed to agree that "in any given proceeding, the . . . standard of proof . . . reflects not only the weight of the private and the public interests affected, but also a societal judgment about how the risk of error should be distributed between the litigants."[24]

## B. The Decision Theoretic Formulas

As many commentators have noted,[25] the notion that the burden of persuasion turns on the relative magnitude of the costs of errors can be recast in the language of decision theory. This translation does more than substitute the bland but exacting terminology of statisticians or similar analysts for the rhetorical flourishes of judicial opinions. It provides a method for arriving at quantitative interpretations of the burdens of persuasion. These quantitative interpretations, which have figured in a few lower court opinions,[26] are in turn important in appreciating why one cannot identify a unique level of "statistical significance" that would correspond to proof satisfying the burden of persuasion appropriate to a given type of case.

A hypothetical, and admittedly contrived, case may help to illustrate the formal, mathematical approach. A gambler represents that certain dice have a 50-50 chance of showing an even or an odd number. He entices the person sitting next to him on a long train trip to play a game in which they alternate rolling the dice and predicting whether the outcome will be an even or odd. The fellow passenger quits after losing thousands of dollars on the sequence,

[19] *Id.* at 371-72 (footnotes omitted).

[20] 441 U.S. 418 (1979).

[21] *Id.* at 423.

[22] 455 U.S. 745 (1982).

[23] The dissenting Justices quoted at length from the concurring opinion in *Winship*, but they argued that, because "the interests at stake are of roughly equal societal importance," a state could adopt the preponderance of the evidence standard without depriving the natural parents of due process. *Id.* at 787.

[24] *Id.* at 755.

[25] *See* Brook, *Inevitable Errors: The Preponderance of the Evidence Standard in Civil Litigation,* 18 TULSA L.J. 79 (1982); Kaplan, *Decision Theory and the Factfinding Process,* 20 STAN. L. REV. 1065 (1968); Kaye, *The Limits of the Preponderance of the Evidence Standard: Justifiably Naked Statistical Evidence and Multiple Causation,* 1982 AM. B. FOUND. RESEARCH J. 487; Lempert, *supra* note 12; Milanich, *Decision Theory and Standards of Proof,* 5 LAW & HUMAN BEHAV. 87 (1981).

[26] *E.g.,* Ethyl Corp. v. EPA, 541 F.2d 1 (D.C. Cir. 1975), *cert. denied,* 426 U.S. 941 (1976); United States v. Fatico, 458 F. Supp. 388 (E.D.N.Y. 1978).

(O,E,O,O,O,O,O,O,O,E). Later, the disgruntled traveller learns that the gambler always carries two sets of dice. One set, the "null dice," are fair, which is to say that they have a probability of 0.5 of producing an odd sum on each toss. The "alternative dice" are loaded so that they have a probability of 0.8 of producing an odd sum each time. The enraged traveller brings an action for deceit. The dice are not available for testing because they have been lost in a fire that destroyed the gambler's home.

Should the jury conclude that the dice were loaded? Considering this question from the perspective of decision theory, the first proposition, clearly accepted by the courts,[27] is that to decide the case is to run a risk of deciding it wrongly. In particular, the jury can make two distinct decisions producing the two types of errors that Justice Harlan described. One possible decision ($D_I$) is to conclude that the dice were the "alternative dice" and return a verdict for plaintiff. The other possibility ($D_o$) is to accept the "null hypothesis" and return a verdict for the defendant. But the jury cannot be certain where the truth lies. If the "null hypothesis" is correct but the jury chooses $D_I$, it registers a false alarm, also called a Type I, or false positive error. If this "Case of the Crooked Gambler" were a criminal proceeding, one could call this decision a false conviction. On the other hand, if the dice were the "alternative dice" and the jury chooses $D_o$, it misses the signal and makes a Type II or false negative error. In the criminal context, one could call this a false acquittal. Using $H_o$ as an abbreviation for the hypothesis that the dice are the null dice (the "null hypothesis") and $H_I$ for the alternative hypothesis, figure 1 summarizes these possibilities.

FIGURE 1
POSSIBLE DECISIONS AND ERRORS

| | | $H_o$ true | $H_I$ true |
|---|---|---|---|
| DECISION BASED ON EVIDENCE | $D_I$: Reject $H_o$ Find for Plaintiff | False Alarm | Correct Verdict |
| | $D_o$: Accept $H_o$ Find for Defendent | Correct Verdict | Miss |

Evaluating all the evidence in the case, the jury makes some rough estimate of the probability that the dice were the alternative dice and the probability that they were the null dice. These probabilities are $Pr(H_I|Evidence)$ and $Pr(H_o|Evidence)$. The "Evidence" in these expressions indicates that these are probabilities conditioned on all the evidence in the case. One need not worry about how the jurors evaluate the evidence to arrive at an estimate of these

[27] *See, e.g.,* In re Winship, 397 U.S. 358, 370 (1970) (Harlan, J., concurring):

> [I]n a judicial proceeding in which there is a dispute about the facts of some earlier event, the factfinder cannot acquire unassailably accurate knowledge of what happened. Instead, all the factfinder can acquire is a belief of what *probably* happened. The intensity of this belief — the degree to which a factfinder is convinced that a given act actually occurred — can, of course, vary. In this regard, a standard of proof represents an attempt to instruct the factfinder concerning the degree of confidence our society thinks he should have in the correctness of factual conclusions for a particular type of adjudication.

probabilities. Although in a real trial the probabilities may never be articulated in quantitative form, one can characterize the end result of this weighing of the evidence by an assertion about the probability or odds in favor of $H_1$. This probability is called a "posterior probability" because it is formed after all the evidence is considered.

Once the posterior probability is available, decision theory dictates the verdict that should be reached—if one can specify the costs of each type of error. Suppose, following what Justice Blackmun once characterized as "perhaps not an unreasonable assumption"[28] for criminal cases, one says that the consequence of an incorrect decision for defendant (a false alarm, or false conviction) is ten times as serious as the consequence of an incorrect decision for plaintiff (a miss, or false acquittal). The rule that should follow for choosing between $D_1$ (a decision for plaintiff's hypothesis) and $D_o$ (a decision for defendant's hypothesis) that minimizes the expected loss[29] is this: decide for plaintiff ($D_1$) if $Pr(H_1|Evidence)$ exceeds $10Pr(Ho|Evidence)$; otherwise, decide for defendant ($D_o$). Thus, the decision theoretic solution is to take action $D_1$ if the posterior odds in favor of $H_1$ are better than ten to one. Expressed as a posterior probability, the standard is $Pr(H_1|Evidence) > 10/11 = 0.91$.

As this example indicates, if the degree of subjective certainty that the law insists upon can be construed as a probability,[30] the burden of persuasion can be understood in quantitative terms as a function of the costs of each type of error. Obviously, there is no single number that could be used to give the relative costs of the two types of error in a criminal case,[31] and the fact that this mathematical structure can be used to explore the effect of different weightings on the numerical expression of the burden of persuasion does not imply that it would be advisable to instruct jurors in numerical terms.[32]

Fortunately, the hypothetical posed here involves a civil case, and it seems plausible to adopt Justice Harlan's suggestion that in the typical civil case where the preponderance of the evidence standard applies, the law treats a mistaken ver-

[28] Ballew v. Georgia, 435 U.S. 223, 234 (separate opinion).

[29] The expected loss under a particular decision rule is (1) the cost of a false alarm weighted by the probability of a false alarm plus (2) the cost of a miss weighted by the probability of a miss. If many cases were decided according to the same decision rule, the mean costs of the errors would approach this expected loss.

[30] The validity of this transformation is not uniformly accepted. *See* Brilmayer & Kornhauser, *Review: Quantitative Methods and Legal Decisions,* 46 U. CHI. L. REV. 116 (1978); Callen, *Notes on a Grand Illusion: Some Limits on the Use of Bayesian Theory in Evidence Law,* 57 IND. L.J. 1 (1982); Cohen, *Subjective Probability and the Paradox of the Gatecrasher,* 1981 ARIZ. ST. L.J. 627. The conceptual value of the subjective interpretation of probability is defended in Kaye, *Paradoxes, Gedanken Experiments and the Burden of Proof: A Response to Dr. Cohen's Reply,* 1981 ARIZ. ST. L.J. 635; Kaye, *The Laws of Probability and the Law of the Land,* 47 U. CHI. L. REV. 34 (1979).

[31] Hale thought that five guilty men should be acquitted before one innocent man was convicted. 2 M. HALE, PLEAS OF THE CROWN 288 (W. Stokes & E. Ingersoll ed. 1847). Blackstone thought the ratio should be ten to one. 4 W. BLACKSTONE, COMMENTARIES 358. Fortescue thought that in capital cases the ratio should be 20 to one. J. FORTESCUE, COMMENDATION OF THE LAWS OF ENGLAND 45 (F. Grignor transl. 1917). *See generally* Fletcher, *Two Kinds of Legal Rules: A Comparative Study of Burden of Persuasion Practices in Criminal Cases,* 77 YALE L.J. 880 (1968); Kaplan, *supra* note 25, at 1077.

[32] *See* Tribe, *supra* note 10. *But see* Nagel, *Bringing the Values of Jurors in Line with the Law,* 63 JUDICATURE 189 (1979) (advocating quantification in instructions).

dict for the plaintiff as neither better nor worse than a mistaken verdict for the defendant.[33] The decision rule which then minimizes the expected costs of mistaken verdicts is to find for plaintiff whenever the posterior odds exceed one, that is, whenever Pr $(H_1|\text{Evidence}) > 0.5$.[34] Under this rule, the judge or jury should proceed on the assumption that the more probable of the two hypotheses is true.

In sum, the decision theoretic interpretation of the applicable legal burden of persuasion is that the stringency of the requirement in different sorts of cases reflects in a rough way the relative costs of false alarms and misses, and that the standards prescribe how the finder of fact should react to the posterior odds. The next step is to see how the standards for establishing facts in scientific inquiry fit into this framework.

## III

## THE SCIENTIFIC STANDARDS OF PROOF

As observed at the outset of this article, there is an understandable tendency on the part of litigants and judges to demand that quantitative scientific evidence meet some unambiguous test for acceptance in the scientific community. The Supreme Court speaks of statistical tests that would make an hypothesis "suspect" to a social scientist,[35] and the lower courts respond by requiring "an objective process known as hypothesis testing."[36]

At the same time, the more sophisticated courts realize that these statistical standards are relatively stringent and often more demanding than the pertinent legal values would dictate. The Court of Appeals for the District of Columbia, sitting en banc in *Ethyl Corporation v. EPA,*[37] expressed this point of view. In upholding the Environmental Protection Agency's determination, based on "conflicting and inconclusive"[38] scientific evidence, that lead emissions from automobiles presented a significant risk to the health of urban populations, the majority rejected the industry's claim that the agency had to rely on a chain of "*scientific* facts for evidence that reputable scientific techniques certify as certain."[39] In doing so, the *Ethyl* court offered an interpretation of the legal burdens of persuasion along the lines described above; however, the court's glancing look at the process for the acceptance of claims in science was less perceptive:

> Typically, a scientist will not so certify evidence unless the probability of error, by standard statistical measurement, is less than 5%. That is, scientific fact is at least 95% certain.
>
> Such certainty has never characterized the judicial or the administrative process. It may be that the 'beyond a reasonable doubt' standard of criminal law demands 95%

[33] *See supra* text accompanying note 19. Not all commentators accept this premise. *See, e.g.,* Orloff & Steadman, *A Framework for Evaluating the Preponderance-of-the-Evidence Standard,* U. PA. L. REV.(1983); Tyree, *Proof and Probability in the Anglo-American Legal System,* 23 JURIMETRICS J. 89 (1982).

[34] *See* Kaye, Book Review, 89 YALE L.J. 601 (1980) (reviewing M. FINKELSTEIN, QUANTITATIVE METHODS IN LAW: STUDIES IN THE APPLICATION OF MATHEMATICAL PROBABILITY AND STATISTICS TO LEGAL PROBLEMS (1978)).

[35] *See supra* text accompanying note 4.

[36] Moultrie v. Martin, 690 F.2d 1078, 1082 (4th Cir. 1982).

[37] 541 F.2d 1 (D.C. Cir. 1975), *cert. denied,* 426 U.S. 941 (1976).

[38] *Id.* at 26.

[39] *Id.* at 28 n.58.

> certainty. . . . But the standard of ordinary civil litigation, a preponderance of the evidence, demands only 51% certainty. . . .
>
> The standard before administrative agencies is no less flexible. Agencies are not limited to scientific fact, to 95% certainties. Rather, they have at least the same factfinding powers as a jury, particularly when, as here, they are engaged in rule making. . . . We must deal with the terminology of law, not science.[40]

The Court of Appeals was on the right track. When scientific studies are relevant, a court or agency must examine the scientific findings with the instruments of legal factfinding. The court's assumption, however, that when the "probability of [statistical] error is less than 5%," the "scientific fact is at least 95% certain" exemplifies a common misunderstanding of the role of statistical tests in scientific inference.[41] To expose this mistake and to elucidate the precise connection between statistical error and satisfying a burden of persuasion, whether it be a requirement of 95% or merely 51% "certainty," an elementary understanding of the mathematical procedure known as hypothesis testing[42] is called for.

The Case of the Crooked Gambler provides an illustration of this procedure. Disregarding some important subtleties, suppose that the experimental results will not be regarded as convincing evidence of the alternative dice unless the probability of obtaining these results by rolling the null dice is no larger than five percent. This requirement is more or less equivalent to insisting on a discrepancy of at least two standard deviations in a case like *Castaneda.*[43] If this is the standard—and it sounds scientific enough—a court cannot conclude that the dice are the alternative ones. The probability of the null dice's producing at least eight out of ten odd outcomes is $\Pr(\text{Evidence}|H_o) = 0.055$, which is slightly more than 5%. In other words, the experimental results are not "statistically significant" at the .05 level. The chances are greater than one in twenty that the dice would exhibit such aberrant behavior even if the dice were exquisitely balanced. Thus, the null hypothesis cannot be rejected with "scientific certainty."

Even if this procedure seems reasonable—and it is not without powerful critics in the world of science[44]—does the finding that the significance probability was just over 5% establish that one should be just shy of 95% certain that the dice were indeed the null dice? If it does, then the hypothesis test outlined above would be

[40] *Id.*

[41] For other opinions advancing substantially the same erroneous interpretation of statistical significance, see, *e.g.,* Moultrie v. Martin, 690 F.2d 1078, 1083 n.7 (4th Cir. 1982); National Lime Ass'n v. EPA, 627 F.2d 416, 453 (D.C. Cir. 1980); United States v. Georgia Power Co., 474 F.2d 906, 915 (5th Cir. 1973). That courts may misconstrue the meanings of "significance" and "confidence" is not surprising. Authors of textbooks and journal articles do the same. *See, e.g.,* D. BARNES, STATISTICS AS PROOF: FUNDAMENTALS OF QUANTITATIVE EVIDENCE 162 (1983); Braun, *Statistics and the Law: Hypothesis Testing and Its Application to Title VII Cases,* 32 HASTINGS L.J. 59, 87 (1980).

[42] The concern here is with what are technically known as Neyman-Pearson test procedures. Other inferential techniques are sometimes loosely referred to as hypothesis or significance tests. *See* V. BARNETT, COMPARATIVE STATISTICAL INFERENCE (2d ed. 1982).

[43] *See, e.g.,* Kaye, Book Review, *supra* note 6. The courts are beginning to appreciate the connection between the significance level (the 5% figure) and the test statistic (the number of standard deviations). *See, e.g.,* EEOC v. Federal Reserve Bank of Richmond, 698 F.2d 633, 654 (4th Cir. 1983). Unfortunately, the *Castaneda* and *Hazelwood* opinions did not make the point with any semblance of clarity. *See* Kaye, *Rejoinder,* 77 J. AM. STATISTICAL A. 790 (1982).

[44] *See, e.g.,* THE SIGNIFICANCE TEST CONTROVERSY: A READER (D. Morrison & R. Henkel eds. 1970).

tantamount to asking for something like proof beyond a reasonable doubt, as the *Ethyl Corporation* court seemed to think. In the language of part I, the hypothesis test would amount to a decision criterion requiring $D_1$ whenever the posterior probability $\Pr(H_1|\text{Evidence})$ exceeded 0.95.

The difficulty is that this interpretation of the result of the hypothesis test is wrong. The test was structured so as to retain the null hypothesis unless the chance of getting the evidence under this hypothesis fell below 5%. The test focused exclusively on the probability of the evidence given the null hypothesis. Nothing was said about the probability of the hypothesis in the light of the experimental evidence. It may be tempting to call the probability of 0.055 the chance of a coincidence, and to say that the probability of something other than a coincidence—of foul play—must be what is left over, namely 0.945. But this only shows that one can "prove" anything with words. The more precise mathematical notation makes it plain that the burden of persuasion refers to one probability—Pr(Alternative hypothesis:Evidence)—while the hypothesis test looks to another—Pr(Evidence:Null hypothesis). There is a well-defined mathematical relationship between these two probabilities, but it is not the simple one that linguistic analysis suggests.[45]

Intuitively, this distinction is not difficult to appreciate. Why, after all, should someone be about 95% confident that the gambler used the alternative dice just because the chance that the null dice would give at least eight out of ten odd numbers is about 5%? The hypothetical did not even consider the probability that so many odd numbers would appear if the alternative dice had been involved. This probability is $\Pr(\text{Evidence}|H_1) = 0.678$. These two probabilities, 0.055 for the evidence given the null hypothesis and 0.678 for the evidence given the alternative hypothesis, summarize all the information that can be found in the statistical data. The probabilities do not change if, instead of a "Crooked Gambler," the person who had proposed the game and won the money was an honest law professor who had inherited the two sets of dice from his distant cousin, the crooked gambler, but who, having no idea of the latter's dishonest inclinations, selected one of the two sets of dice at random.

In short, if both the null and alternative hypotheses were equally likely at the outset, as they presumably would be in the case of the Honest Law Professor, one could reasonably conclude from the statistical evidence that the probability in favor of the alternative dice really is in the vicinity of 0.95. But the assumption that the gambler is just as likely to pull out the alternative dice as the null dice has no evidentiary foundation. His behavior would seem to depend on such factors as his avarice and his perception of the gullibility of the plaintiff. Although such matters, which might be summed up in a statement about the "prior probability" that the gambler would use the null dice, surely bear on the probability that the gambler used the alternative dice to part the plaintiff from his money, they are outside the scope of the hypothesis test. That test looks solely to Pr(Evidence|Null hypothesis). The test ignores the two other vital ingredients, the probability of the

[45] For the precise relationship, *see, e.g.,* M. DEGROOT, PROBABILITY AND STATISTICS 373-82 (1975).

evidence given the alternative hypothesis and the prior probability of the hypothesis itself.[46]

For these reasons, the results of hypothesis tests do not dictate "legal significance." They do not even determine the degree of certainty that a scientific hypothesis enjoys. The significance probability Pr(Evidence|Null hypothesis) used in statistical tests, whether stated explicitly or measured indirectly with a statistic like the standard deviation, is but part of the equation. It does not necessarily mean that the posterior odds exceed the threshold implicit in the pertinent burden of persuasion. The finder of fact also needs to know the prior probability and the probability of the evidence under the alternative hypothesis.[47] Expert testimony sometimes can quantify the latter probability, but the former is outside the range of statistical expertise.[48]

## IV

## CONCLUSION

Statistically significant results are nice to have. Scientists like them, and now litigants who rely on statistical evidence also want them. But the mere fact that an expert states that data are "significant" does not necessarily mean that the evidence satisfies the applicable burden of persuasion. Nor does the fact that a scientist cannot certify data as "significant" imply that the evidence inevitably falls short of what the law requires. The process of judgment, in law as well as in science, is much richer than the recipes for statistical hypothesis tests reveal.

[46] One commentator, recognizing that the 0.05 significance level often is more demanding than the preponderance of the evidence standard, proposes structuring hypothesis tests to equate the risks of false alarms and misses. Dawson, *Are Statisticians Being Fair to Discrimination Plaintiffs?*, 21 JURIMETRICS J. 1 (1980). This procedure, however, does not conform to the preponderance of the evidence standard either, unless the prior odds happen to be 50-50.

[47] Because, roughly speaking, only one of the three pertinent variables is known from the significance probability Pr(Evidence | Null hypothesis), a small value for this variable (that is, a "highly significant" result), does not guarantee a large value for the posterior probability Pr(Alternative hypothesis | Evidence). For an example drawn from science, see Kaye, Book Review, 32 J. LEGAL EDUC. 145, 149 (1982) (reviewing P. SHUCHMAN, PROBLEMS OF KNOWLEDGE IN LEGAL SCHOLARSHIP (1979)). Inversely, a large significance probability (a "very insignificant result") need not imply a small posterior probability. For instance, if only six of ten rolls of the dice had produced odd numbers in the Crooked Gambler case, the probability of such evidence under the null hypothesis would be 0.377, which is "not significant." Yet, the probability under the alternative hypothesis is also much higher, specifically 0.967, so that if the prior odds were more or less equal, the statistical evidence would favor the alternative hypothesis.

[48] *See* Aickin & Kaye, *Some Mathematical and Legal Considerations in Using Serological Tests to Prove Paternity*, in INCLUSION PROBABILITIES IN PARENTAGE TESTING (R. Walker ed. 1983); Ellman & Kaye, *Probabilities and Proof: Can HLA and Blood Group Testing Prove Paternity*, 54 N.Y.U. L. REV. 1131 (1979).