Washington Law Review

Volume 61
Number 4 *Dedicated to Robert Meisenholder*

10-1-1986

Is Proof of Statistical Significance Relevant?

D.H. Kaye

IS PROOF OF STATISTICAL SIGNIFICANCE RELEVANT?

D.H. Kaye*

In the Old Testament it is written that "Varying weights, varying measures, are both an abomination to the Lord."[1] In the classic treatises on evidence it is written that the court or jury must weigh the evidence, and upon weighing it, determine whether the plaintiff or the defendant prevails. In assessing most evidence, courts are comfortable with the lack of an orthodox set of weights and measures. However, some courts have indicated that statistical evidence may well be cast out—if not as an abomination, as a scientific charlatan—unless it is subjected to a procedure known as "hypothesis testing."[2] Roughly speaking, a hypothesis or significance test determines whether an observed result is so unlikely to have occurred by chance alone that it is reasonable to attribute the result to something else. There are many rather mechanical procedures for performing these tests and a number of judges, attorneys, and law professors have suggested that hypothesis testing provides an objective, scientific means of settling disputed questions on which statistical evidence is brought to bear.[3] Discrimination litigation, environmental cases, food and drug regulation, and a variety of other administrative and judicial proceedings are obvious arenas for hypothesis testing.[4] Differences between the percentage of blacks selected for grand juries and the percentage in the community,[5]

* Professor of Law and Director, Center for the Study of Law, Science and Technology, Arizona State University. The author is indebted to Hans Zeisel for commenting on a draft of the article and to Mikel Aickin for his insights into the role of statistical analysis in litigation.

1. *Proverbs* 20:10.

2. There are various types of "hypothesis testing." Neyman-Pearson testing, which is the most common and the main concern here, is conceptually distinct from Bayesian hypothesis tests. *See* M. DEGROOT, PROBABILITY AND STATISTICS 381 (1975). The usefulness of Bayes test procedures for forensic purposes is questioned in Kaye, *Hypothesis Testing in the Courtroom*, in CONTRIBUTIONS TO THE THEORY AND APPLICATION OF STATISTICS (A. Gelfand ed. in press). In addition, although I shall use the phrases "hypothesis testing" and "significance testing" interchangeably, one can distinguish between them. *See infra* note 107.

3. *See, e.g.*, Moultrie v. Martin, 690 F.2d 1078, 1082 (4th Cir. 1982); Braun, *Statistics and the Law: Hypothesis Testing and Its Application to Title VII Cases*, 32 HASTINGS L.J. 59, 87 (1980).

4. *See generally* D. BARNES, STATISTICS AS PROOF: FUNDAMENTALS OF QUANTITATIVE EVIDENCE (1983); C. CLEARY, MCCORMICK ON EVIDENCE §§ 208–211 (3d ed. 1984) [hereinafter MCCORMICK]; Curtis & Wilson, *The Use of Statistics and Statisticians in the Litigation Process*, 20 JURIMETRICS J. 109 (1979).

5. *E.g.*, Vasquez v. Hillery, 106 S. Ct. 617 (1986); Boykins v. Maggio, 715 F.2d 995 (5th Cir. 1983), *cert. denied*, 466 U.S. 940 (1984). *See generally* Kaye, *Statistical Analysis in Jury Discrimination Cases*, 25 JURIMETRICS J. 274 (1985).

between the wages[6] or promotions[7] of male and female employees, between the rates at which blacks and whites found guilty of capital offenses are sentenced to death,[8] between the incidence of asbestosis in workers exposed to high levels of asbestos dust as opposed to nonexposed workers,[9] and between the rates of cancers among rats fed large amounts of the food coloring red dye number two as opposed to a control group of rats[10] exemplify the many cases in which courts or administrators have puzzled over the meaning of hypothesis tests. Considering the frequently voiced suspicion that statistics can prove anything,[11] an unvarying set of weights and measures for statistical evidence would be a welcome antidote to more nefarious or less sophisticated presentations.

This article examines the status of significance testing in litigation. Part I describes the case law on the need for the procedure. Part II explains the nature and terminology of hypothesis testing as used in court. Part III enumerates some of the problems that arise in these forensic applications, and Part IV pursues one such problem—that of selecting a "significance level." These sections show that explicit hypothesis testing is poorly suited for courtroom use. Statements as to what results are or are not "statistically significant" should be inadmissible. Part V suggests the use of other statistical tools and terms that do not "test" hypotheses but can better aid the finder of fact in judging the probative value of the statistical evidence.

I. THE DEMAND FOR HYPOTHESIS TESTING IN THE COURTROOM

The idea that formal hypothesis tests should or must be used to assist the judge or jury in evaluating statistical evidence is a recent phenomenon. Before 1970, almost no federal cases adverted to "statistically significant" evidence.[12] In the early seventies, a trickle of reported cases mentioned significance tests. Then, in 1977, an event that only attorneys could call dramatic happened. The United States Supreme Court calculated a statistic

6. *E.g.*, Valentino v. United States Postal Serv., 674 F.2d 56, 70–71 (D.C. Cir. 1982).

7. *E.g.*, Sainte Marie v. Eastern R.R. Ass'n, 650 F.2d 395 (2d Cir. 1981).

8. McCleskey v. Zant, 580 F. Supp. 338 (N.D. Ga. 1984), *rev'd on other grounds en banc sub nom.* McClesky v. Kemp, 753 F.2d 877 (11th Cir. 1985) *cert. granted in part*, 106 S. Ct. 331 (1986).

9. Reserve Mining Co. v. Environmental Protection Agency, 514 F.2d 492 (8th Cir. 1975).

10. Certified Color Mfrs. Ass'n v. Mathews, 543 F.2d 284 (D.C. Cir. 1976).

11. *E.g.*, EEOC v. Federal Reserve Bank, 698 F.2d 633, 645–46 (4th Cir. 1983), *rev'd on other grounds sub nom.* Cooper v. Federal Reserve Bank, 467 U.S. 867 (1984).

12. A search on December 11, 1984 of the general federal library of the LEXIS database revealed that 519 cases used the words "statistically significant" or "statistical significance." Nearly two-thirds of these cases were decided in the past four years, and only seven—barely more than one percent—were dated before 1970.

known as the standard deviation.[13] In footnotes to two opinions, *Castaneda v. Partida,*[14] and *Hazelwood School District v. United States,*[15] the Court not only performed a few textbook calculations, but it also spoke of "two or three standard deviations" as being necessary to establish statistical significance.[16] The lower courts reacted. In the following year, nearly forty published opinions discussed the statistical sigificance of numerical evidence. Although the Supreme Court had stated that its computations were not intended to imply that this procedure always should be followed,[17] in *Moultrie v. Martin*[18] the Court of Appeals for the Fourth Circuit held that "in all cases involving racial discrimination, the courts of this circuit must apply a standard deviation analysis such as that approved by the Supreme Court in *Hazelwood* before drawing conclusions from statistical comparisons."[19] The court reasoned that:

> When a litigant seeks to prove his point exclusively through the use of statistics, he is borrowing the principles of another discipline, mathematics . . . . [He] cannot be selective in which principles are applied. He must employ a standard mathematical analysis. Any other requirement defies logic to the point of being unjust. Statisticians do not simply look at two statistics . . . and make a subjective conclusion that the statistics are significantly different. Rather, statisticians compare figures through an objective process known as hypothesis testing.[20]

While no other circuit appears to have gone to this extreme, most discrimination plaintiffs relying on statistical evidence prize figures that are "statistically significant," and most defendants are delighted if they can demonstrate that the numbers are "not statistically significant." Thus, many lower courts in Title VII cases have come to expect a "standard deviation analysis" and to regard quantitative proof not couched in these terms with suspicion, if not hostility.[21] In these jurisdictions, hypothesis

13. The standard deviation measures the variability, or dispersion, of a batch of numbers. If all the numbers are the same, the standard deviation is zero. If many of the numbers are far from the mean for the entire set, the standard deviation is large.

14. 430 U.S. 482 (1977) (grand jury discrimination).

15. 433 U.S. 299 (1977) (racial discrimination in employment).

16. *Hazelwood,* 433 U.S. at 311 n.17; *Casteneda,* 430 U.S. at 496 n.17. For criticism of the statistical analysis in *Hazelwood,* see, for example, Kaye, *Book Review,* 80 MICH. L. REV. 833, 838–41 (1982); Smith & Abram, *Quantitative Analysis and Proof of Employment Discrimination,* 1981 U. ILL. L. REV. 33, 52–53.

17. *Hazelwood,* 433 U.S. at 311 n.17.

18. 690 F.2d 1078 (4th Cir. 1982).

19. *Moultrie,* 690 F.2d at 1082.

20. *Id.*

21. *See, e.g.,* Hill v. K-Mart Corp., 699 F.2d 776, 780 (5th Cir. 1983). The details of the "standard deviation analysis" are not important to this article, but one cannot help observing that the apparent

testing has become a practical necessity in cases involving statistical proof.[22]

Despite the prevalence of hypothesis testing in discrimination litigation, cases involving scientific identification evidence rarely mention the statistical significance of the forensic scientist's findings. For example, in *State* ex rel. *Hausner v. Blackman*,[23] the Kansas Supreme Court held that it was error in a paternity action to allow testimony that blood group tests on the mother, child, and alleged father implicated the defendant and that the probability that these tests would exculpate a falsely accused man was .70.[24] In other words, the testimony that the court held inadmissible was that the probability that the mother, child, and defendant would have had the blood group types they did was .30 if the defendant was not the biological father. Although the details of the "standard deviation analysis" have no bearing here, the perspective of hypothesis testing which underlies that analysis implies that the blood group data in *Hausner* do not support (to the degree required in *Castaneda* and *Hazelwood*) the hypothesis that the defendant is not the biological father.[25]

In other cases, experts have testified to much smaller probabilities that would make a claim "suspect" in the manner described in *Castaneda* and *Hazelwood*. Once again, however, neither the experts nor the courts have employed these infinitesimal probabilities for significance testing. In *Commonwealth v. Drayton*,[26] for instance, a fingerprint expert stated that the probability that the fingerprints of two different persons would match on

---

infatuation of the courts with this one procedure is such that, all too often, it is employed to the exclusion of other, more appropriate methods. Kaye, *Ruminations on Jurimetrics: Hypergeometric Confusion in the Fourth Circuit*, 26 JURIMETRICS J. 215 (1986); Meier, Sacks & Zabell, *What Happened in Hazelwood: Statistics, Employment Discrimination and the 80% Rule*, 1984 AM. B. FOUND. RES. J. 139; Sugrue & Fairley, *A Case of Unexamined Assumptions: The Use and Misuse of the Statistical Analysis of* Castaneda/Hazelwood *in Discrimination Litigation*, 24 B.C.L. REV. 925 (1983).

22. In the words of one participant, "[t]he judges don't understand how far away three standard deviations is from two but they have finally set out a rule. . . . You see complaint after complaint filed in federal district court mentioning two standard deviations." Michelson, *Statistical Determination in Employment Discrimination Issues*, in THE USE/NONUSE/MISUSE OF APPLIED SOCIAL RESEARCH IN THE COURTS 109, 111–12 (M. Saks & C. Baron eds. 1980).

23. 233 Kan. 223, 662 P.2d 1183 (1983), *aff'g*, 7 Kan. App. 2d 693, 648 P.2d 249 (1982).

24. The court seemed to hold that evidence of the failure to exclude the defendant was inadmissible. *Hausner*, 662 P.2d at 1190. The court also complained that the testimony as to the probability of this outcome was entirely irrelevant to the determination of paternity. *Id.* at 1188. This is patently fallacious, but perhaps the court meant that the expert did not explain the calculation in a way that would have made it sufficiently useful to the jury.

25. *See, e.g.*, Aickin & Kaye, *Some Mathematical and Legal Considerations in Using Serological Tests to Prove Paternity*, in INCLUSION PROBABILITIES IN PARENTAGE TESTING 155 (R. Walker ed. 1983). There may be a subtle fallacy in defining the null hypothesis in this fashion. *See* Aickin, *Some Fallacies in the Computation of Paternity Probabilities*, 36 AM. J. HUMAN GENETICS 904, 907–08 (1984).

26. 386 Mass. 39, 434 N.E.2d 997 (1982).

twelve points of comparison was "one out of 387 trillion."[27] Although the Supreme Judicial Court of Massachusetts held that the expert had no adequate basis for testifying to this probability,[28] even when there is a solid empirical foundation for calculation, most courts will admit the testimony without considering its "statistical significance."[29]

It seems difficult to justify this difference in the treatment of evidence apparently amenable to statistical analysis. If hypothesis testing is the preferred way to evaluate statistical evidence of discrimination, then in the absence of some cogent reason to think otherwise, hypothesis testing should also be the method of choice for assessing identity evidence. Thus, the growing insistence and reliance on hypothesis testing raise both doctrinal and practical problems.

The purpose of expert statistical testimony is to assist the trier of fact in evaluating numerical information. Judges and juries must resolve disputed factual questions as best they can, and they should not delegate this decisionmaking task to statisticians, economists, social scientists, and other experts by trusting to superficially impressive methods whose seeming objectivity does not withstand analysis. "Hypothesis testing" is a technical term for procedures that have important limitations, and "statistical significance" is a phrase that is easily misunderstood. Before any general requirement for employing statistical test procedures evolves out of practice or pronouncement, the nature of hypothesis testing and its limitations and possible disadvantages in forensic applications should be clearly understood. Part II offers an elementary explanation of the ideas underlying significance testing as a preliminary step in elucidating some of the problems with hypothesis tests as devices for evaluating statistical evidence.

II. THE LOGIC OF HYPOTHESIS TESTING

The essential idea behind the concept of statistical significance is easily grasped. To introduce some of the terminology and steps involved in performing a significance test, we shall consider a situation loosely based

27. *Drayton*, 434 N.E.2d at 1005.

28. *Id.* at 1005–06. A depressing number of cases in which probabilites computed without an adequate empirical foundation have been bandied about in court are collected in MCCORMICK, *supra* note 4, § 210. The most notorious is People v. Collins, 68 Cal. 2d 319, 438 P.2d 33, 66 Cal. Rptr. 497 (1968). For a thoughtful and sophisticated analysis of a much earlier case, see Meier & Zabell, *Benjamin Peirce and the Howland Will*, 75 J. AM. STATISTICAL ASS'N 497 (1980).

29. *See* MCCORMICK, *supra* note 4, § 210. The exception is State v. Carlson, 267 N.W.2d 170 (Minn. 1978), and its progeny. In Braun, *Quantitative Analysis and the Law: Probability Theory as a Tool of Evidence in Criminal Trials*, 1982 UTAH L. REV. 41, the author argues for increased use of probablity calculations.

on the facts in *Moultrie v. Martin,* the case in which the Fourth Circuit announced its hypothesis testing requirement.[30] A black defendant, convicted in 1977 in South Carolina of murdering a deputy sheriff, wishes to obtain a writ of habeas corpus from federal court on the theory that the grand jury that indicted him was selected in a way that discriminated against blacks. He consults an attorney who discovers that in South Carolina, jury commissioners examine voter registration lists (which reveal the race of the voters) to prepare a list of persons eligible for grand jury service. To illustrate the simplest sort of hypothesis test, let us pretend that in 1977, the commissioners, intent on discriminating against blacks, prepared two such lists. One, which we shall call the "null list," is perfectly representative of the voting list. Thirty-eight percent of the voters are black, and thirty-eight percent of the persons on the null list are black. The other list, which the officials kept secret and which we shall call the "alternative list," is only fifteen percent black. In 1977, the commissioners selected eighteen persons from one of these lists to serve on the grand jury. Three of these grand jurors, or seventeen percent, were black, and fifteen were white. The commissioners insist that although they had prepared the alternative list as part of a plan to prevent there being "too many" black grand jurors, they abandoned this plan and used only the official, null list.

Petitioner's counsel believes that in view of the existence of the secret list, the disparity between the proportion of blacks on the voting list (.38) and the proportion on the grand jury (.17) supports the claim that the commissioners illegally drew the grand jury from the alternative list. However, since counsel has heard that the appellate courts are beginning to insist on "statistically significant" disparities, she warns her client that if he does not come forward with the results of "an objective process known as hypothesis testing," he may lose his case.

At this point, a statistical consultant enters the case. He sets up a statistical test to choose between two hypotheses. The first hypothesis he calls the "null hypothesis," and he writes it like this:

$$H_0: \theta = .38$$

When counsel (and later the judge) asks the expert what this means, he says that H_0 is an abbreviation for "null hypothesis," and that the Greek letter θ (theta) is a "parameter." Here, θ is the probability of selecting a black juror on each independent draw from the list. The value of θ is unknown, but the null hypothesis asserts that it is .38, which is to say that the null list was used. To be understandable, the consultant offers an analogy. He suggests that one should think of the null hypothesis as an assertion like "the

30. *See supra* text accompanying note 18. Later we shall modify the more fanciful features of this example to provide a more accurate and complete renditon of the actual facts in *Moultrie.*

defendant is not guilty." The hypothesis test is like a criminal trial that will accept the null hypothesis H_0 unless there is sufficient statistical evidence against it.[31]

The consultant next identifies an "alternative hypothesis," which he writes as H_1: $\theta = .15$. This, he suggests, is like the government's claim that the defendant is guilty. H_1 asserts that the commissioners resorted to the secret list from which a black has only a fifteen percent chance of being chosen on each independent draw.

Now for the statistical test. The consultant computes a "P-value." Roughly speaking, he says, this is the probability of obtaining the observed disparity (or an even greater disparity) if the null list had been used. In symbols, P-value = Pr(Extreme Data$|H_0$). Leafing through a book and muttering something about interpolating from a table of binomial probabilities, the consultant says that the P-value for this data is .051. This, he concludes, is not good enough to be "statistically significant" at the .05 level.[32] In other words, the chances are greater than one in twenty that the random sampling from the null list would produce a grand jury with no more than three blacks. Therefore, the null hypothesis cannot be rejected. Petitioner loses. Or does he?

At first glance, it might seem that the statistical analysis has demonstrated that petitioner's evidence is too weak to make out even a prima facie case of racial discrimination.[33] The statistician's conclusion that the small number of black jurors is not "significant" is the result of an objective procedure—in the sense that anyone who correctly follows the unambiguous steps will come to the same conclusion. But this objectivity begs the question. The real issue for the law is not whether every expert who follows the same recipe will agree that the observations are not "significant" at the .05 level. Rather, two evidentiary issues are present. With regard to the weight of the finding, the pertinent question is whether such uncontroverted testimony dictates the presence or absence of a prima facie case. As to the finding's admissibility, the issue is whether the testimony that the numbers are "significant" sufficiently advances the understanding

31. Reliance on this analogy is not entirely hypothetical. *See, e.g.*, D. BARNES, *supra* note 4, at 146; Feinberg, *Teaching the Type I and Type II Errors: The Judicial Process*, AM. STATISTICIAN, June 1971, at 30. It can be misleading, however, because the significance level bears no simple relationship to the the burden of persuasion. Kaye, *supra* note 2; Kaye, *Statistical Significance and the Burden of Persuasion*, 46 LAW & CONTEMP. PROBS. 13 (Autumn 1983).

32. Part III.*A* discusses the ubiquitous .05 level.

33. On the role of statistics in establishing a prima facie case in discrimination litigation, see generally Segar v. Smith, 738 F.2d 1249 (D.C. Cir. 1984); D. BALDUS & J. COLE, STATISTICAL PROOF OF DISCRIMINATION (1980); W. CONNOLLY & D. PETERSON, USE OF STATISTICS IN EQUAL EMPLOYMENT OPPORTUNITY LITIGATION (1980). The Supreme Court has implied that a P-value of .05 or less is needed to establish a prima facie case of disparate treatment. Casteneda v. Partida, 430 U.S. 482, 497 n. 17 (1977). *See generally* Kaye, *supra* note 16; *infra* note 51.

of the trier of fact to be worth the effort consumed in its presentation and explanation.

Before confronting these questions, however, it is worth stating how the perspective that our imaginary statistical consultant adopted captures the essence of hypothesis testing even with more esoteric statistical models. In our *Moultrie* variation, a simple model of the process giving rise to the data enabled the consultant to perform the hypothesis test. The consultant posited that each draw from a voter list was independent with a fixed, but unknown, probability (depending on which list was used) of producing a black juror. This picture of the selection process is a probability model. The unknown probability—technically called the parameter of the model—was either .38 (if the null list had been used) or .15 (if the alternative list had been employed). The hypothesis test used here focused on the particular value of θ in the context of this model. Distinct values for θ make certain outcomes more likely than others, and the probability of various extreme outcomes arising when θ has the value given by the null hypothesis is the P-value.

The same concepts underlie hypothesis testing of parameters of the more complex models that are becoming familiar in discrimination litigation,[34] in antitrust cases,[35] in estimating lost profits,[36] and in certain administrative proceedings.[37] The statistical models typically involve parameters whose values are unknown.[38] Data from records such as employee files can be used to estimate the values of these parameters. The theory behind hypothesis testing in such settings is that if the model were to generate not one batch of data, but repeated batches, the values for the parameters estimated from each batch of data would be distributed about the true value in a probabilistically well-defined way. Knowledge of this theoretical distribution of the estimates about the true value leads to the P-value.[39]

34. *E.g.*, Lehman v. Trout, 465 U.S. 1056 (1984); Valentino v. United States Postal Serv., 511 F. Supp. 917, 944 (D.D.C. 1981), *aff'd*, 674 F.2d 56 (D.C. Cir. 1982); Presseisen v. Swarthmore College, 442 F. Supp. 593 (E.D. Pa. 1977) *aff'd* mem., 582 F.2d 1275 (3rd Cir. 1978); Rubinfeld, *Econometrics in the Courtroom*, 85 COLUM. L. REV. 1048 (1985). *Cf.* Coble v. Hot Springs School Dist., 682 F.2d 721, 730–33 (8th Cir. 1982) (chiding plaintiffs for not applying multiple regression analysis).

35. *See, e.g.*, Finkelstein & Levenbach, *Regression Estimates of Damages in Price-Fixing Cases*, 46 LAW & CONTEMP. PROBS. 145 (Autumn 1983); Rubinfeld & Steiner, *Quantitative Methods in Antitrust Litigation*, 46 LAW & CONTEMP. PROBS. 69 (Autumn 1983).

36. *E.g.*, Christian Broadcasting Network v. Copyright Royalty Tribunal, 720 F.2d 1295 (D.C. Cir. 1983), *cert. denied*, 106 S. Ct. 1245 (1986); Spray-Rite Serv. Corp. v. Monsanto Co., 684 F.2d 1226 (7th Cir. 1982), *aff'd*, 465 U.S. 752 (1984).

37. *See, e.g.*, South Dakota Pub. Util. Comm'n v. Federal Energy Regulatory Comm'n, 643 F.2d 504, 513 n.13 (8th Cir. 1981); Finkelstein, *Regression Models in Administrative Proceedings*, 86 HARV. L. REV. 1442 (1973).

38. Nonparametric methods exist, but they do not appear to be used very often in litigation.

39. Processes such as salary assignments or promotions do not always lend themselves to convincing stochastic models. For a way to interpret P-values in these situations, see Freedman & Lane,

For example, in *Segar v. Smith,*[40] black employees of the Drug Enforcement Administration (DEA) alleged that the DEA discriminated against its black agents in salaries, promotions, and other matters. Plaintiffs hired economists to develop a linear regression model relating salaries of DEA agents to years of federal experience,[41] years of nonfederal experience, education, and race. That is, the experts posited that the salary each DEA agent receives can be described by an equation that involves: (a) a coefficient times the number of years of employment with the federal government; (b) another coefficient times the years of nonfederal experience; (c) a third coefficient times some measure of educational attainment (the opinion does not describe this variable); (d) a fourth coefficient times the race of the agent;[42] and (e) an error term with certain convenient statistical properties. The four coefficients—including the coefficient of the race variable—are unknown parameters. For brevity, let us call the coefficient for race by the Greek letter β (beta). The regression analysis uses the records of the employees' salaries to estimate β. Derived from a particular batch of records, this estimate is called a statistic to distinguish it from the parameter that it estimates. In *Segar,* for employees hired after 1972 and on the payroll in October 1978, the estimated value of β was -$1,026. Assuming, among other things, that there is neither interaction nor correlation between race and the other variables that determine salary, this statistic indicates that, on average, a black agent received about a thousand dollars less than a white agent of the same experience and education. But -$1,026 is only an estimate based on the data at hand. If there were a different group of employees, and hence different data, the estimated value of β might depart from -$1,026.

Recognizing this variability in the statistic that estimates β, the *Segar* experts tested whether the coefficient of -$1,026 was significantly different from zero. They took the null hypothesis to be that the parameter for race is zero. In symbols, H_0: $\beta = 0$. If this null hypothesis, and the assumptions listed above are correct, then an agent's race would have no impact on the salary he or she received. A black and a white agent who are equal with respect to all other variables would receive the same salary (subject to an amount given by the error term that reflects the inherent variability in setting salaries and the analyst's inability to take into account every factor

Significance Testing in a Nonstochastic Setting, in A FESTSCHRIFT FOR ERICH L. LEHMAN 185 (P. Bickel, K. Doksum & L. Hodges, Jr. eds. 1983).

40. 738 F.2d 1249 (D.C. Cir. 1984), *cert. denied,* 105 S. Ct. 2357 (1985).

41. The court of appeals stated that the variable was "prior federal experience," *Segar,* 738 F.2d at 1261, while the district court wrote that the variable was "years of federal experience." Segar v. Civiletti, 508 F. Supp. 690, 696 (D.D.C. 1981). Neither opinion gives a full description of the fitted equation.

42. Presumably, race is coded as a one if the agent is black and a zero otherwise.

that determines salaries). An alternative hypothesis is that the coefficient for race is different from zero, that is, H_1: $\beta \neq 0$. If certain additional restrictive assumptions about the error term hold, then the analyst can compute the probability that the estimated value for β would be at least as far from zero as it turned out to be if the null hypothesis were true. If we can be a little bit loose with the term "data,"[43] this probability can be abbreviated as Pr(Extreme Data$|H_0$)—the probability of finding the data (or other data that are no more supportive of H_0) given that the null hypothesis H_0 is correct. In other words, this probability is the P-value for the estimated race coefficient. In *Segar v. Smith,* this number was less than .05; hence, plaintiffs' experts reported that race was "significant" at the .05 level. The court of appeals, reviewing such results, concluded that the regression analysis had "uncovered evidence of significant discrimination in salary levels. . . ."[44]

Segar and our variation on *Moultrie* convey some sense of how hypothesis tests are used in court. The details of the tests will vary.[45] Standard deviations may be mentioned in one case, but not in another.[46] Still, the logic of statistical significance does not change. The statistician posits a probabilistic model of the process giving rise to the data. This model may be a simple binomial model, as in *Moultrie,* a more involved regression model, as in *Segar,* or it may be something else entirely. Whatever it is, it has unknown parameters, and the hypothesis test is supposed to say something about these parameters. The statistician computes the probability that the model will generate data at least as aberrational as the observed data if the value of the parameter specified by the "null hypothesis" is true. If this probability is below .05, the statistician concludes that the observed data are "statistically significant" evidence that the unknown parameter has the value stated in the "alternative hypothesis."

III. THE LIMITATIONS OF HYPOTHESIS TESTING

A. *Selecting a Significance Level*

The forensic applications of hypothesis tests presented in Part II are explicit about the P-value needed for "significant" results. Careful and honest experts will explain that significance has (or has not) been found at a particular level, such as .05. They will say that one can (or cannot) reject the null hypothesis *at this significance level.* Unfortunately, not all experts

43. The calculated coefficient, like any other statistic, is a function of the data.
44. *Segar,* 738 F.2d at 1263.
45. *See supra* note 21.
46. *Id.*

are this precise, and the courts have been impressed with such conclusory statements as "a variance [sic] in excess of 2.33 standard deviations is a 'highly statistically significant disparity.'"[47]

Where the experts are clear about a significance level, they, like *Segar's* economists, tend to choose the .05 level. Presumably, they adopt this figure because it sees frequent use in many academic fields. While recognizing that "the law has not set any precise level at which statistical significance can be said to be sufficient to permit an inference of discrimination,"[48] the court of appeals in *Segar* found various statistical showings to "support an inference" when the .05 level was satisfied and "not to permit an inference" when this level was not attained.[49] The only reason given for the .05 level was that "social scientists usually accept a study that achieves statistical significance at the .05 level."[50] In this regard, the *Segar* court was following the lead of the Supreme Court, which previously had pointed to the popularity of this number among social scientists.[51]

This reverance for social scientific norms may be encouraging to some social scientists, but it should prompt us to ask why the .05 figure has achieved such prominence in that domain. Social scientists did not devise most of the statistical methods seen in court, they did not originate hypothesis testing, and they did not establish the .05 level as anything special. Rather, social scientists adopted the methods and conventions of others who were concerned primarily with problems in biology. The practice of using certain standard levels of significance, particularly .05, can be traced to the influence of the eminent British statistician Sir R.A. Fisher.[52] Fisher wrote:

47. Harrell v. Northern Elec. Co., 672 F.2d 444, 446–47 (5th Cir. 1982); *cf.* Lewis v. NLRB, 750 F.2d 1266, 1272 (5th Cir. 1985) (court refers to "statistically significant" results without stating the significance level of the P-value); Miles v. M.N.C. Corp., 750 F.2d 867, 873 (11th Cir. 1985) (same).

48. *Segar*, 738 F.2d at 1282.

49. *Id.* at 1283. Relying on a finding of the district court, the court of appeals suggested that the reason that certain statistics did not achieve acceptable levels of statistical significance was not that the null hypothesis was true, but rather that the sample size "was too small to generate statistically significant evidence of discrimination . . . ." *Id.* Putting the weak statistical showing to one side, the court of appeals held that enough other probative evidence existed to support the district court's determination that the DEA discriminated in promotions. *Id.*; *cf.* Coser v. Moore, 739 F.2d 746, 754 n.3 (2d Cir. 1984) ("While recognizing that [the .05] significance level has no talismanic importance, we accept it for purposes of this case as a measure of validity.").

50. *Segar*, 738 F.2d at 1282. The court referred also to the fact that the Justice Department's Uniform Guidelines on Employee Selection rely on the .05 level. *Id.* at 1282–83.

51. In Castaneda v. Partida, 430 U.S. 482, 497 n.17 (1977), the Court observed that a disparity of two or three standard deviations would be "suspect" to a social scientist. The P-value for a disparity of two standard deviations in either direction from the mean of a normally distributed random variable is about .05.

52. Fisher, a statistician and geneticist at the agricultural experiment station at Rothamsted, England, was the father of the randomized experiment, the general use of regression, and the mathematical derivation of the probability distributions of several important test statistics. He was not

> It is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." . . . If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach that level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance.[53]

As one contemporary statistician has remarked: "There you have it. Fisher thought 5% was about right, and who was there to disagree with the master?"[54]

As Fisher's explanation reveals, there is no sharp border between "significant," and "insignificant." Although a few commentators and courts have inadvertently suggested otherwise,[55] as the P-value decreases, evidence gradually becomes stronger.[56] As a result, most modern statistics texts and journals discourage the reporting of results as "significant" or "insignificant" in favor of explicit statements of P-values. Courts should do likewise. There is no strictly objective basis, in science or in anything else,

---

the originator of tests of significance, but his writings on statistics in scientific research were exceedingly influential.

53. Fisher, *The Arrangement of Field Experiments*, 33 J. MINISTRY AGRIC. GR. BRIT. 504 (1926), as quoted in Savage, *On Rereading R.A. Fisher*, 4 ANNALS OF STATISTICS 471 (1976). Despite this quotation, Fisher did not simply report results as "significant" or "not significant." He made liberal use of P-values in his work, and he cautioned his fellow statisticians that "[w]e have the duty of formulating, of summarising, and of communicating our conclusions, in intelligible form, in recognition of the right of *other* free minds to utilize them in making *their own* decisions." Fisher, *Statistical Methods and Scientific Induction*, 17 J. ROYAL STATISTICAL SOC'Y SERIES B 69, 77 (1955).

54. D. MOORE, STATISTICS: CONCEPTS AND CONTROVERSIES 292 (1979).

55. *E.g.*, Watkins v. Scott Paper Co., 6 Empl. Prac. Dec. (CCH) 8912 (S.D. Ala. 1973) ("If chi-squared or phi reaches a certain level for a certain sample size, validity is established."); Delgado, *Beyond* Sindell: *Relaxation of Cause-In-Fact Rules for Indeterminate Plaintiffs*, 70 CALIF. L. REV. 881, 885 n.19 (1982) ("If [the number of cases of a disease corresponding to the significance level] is represented by 100 + N, then cases beyond this number are evidence of a new cause or agent"); Sperlich & Jaspovice, *Methods for the Analysis of Jury Panel Selections: Testing for Discrimination In a Series of Panels*, 6 HASTINGS CONST. L.Q. 787, 794 (1979) ("probabilities fall into two classes: significant and nonsignificant"); Note, *Statistics as Evidence of Age Discrimination*, 32 HASTINGS L.J. 1347, 1354 (1981) ("The rejection of the null hypothesis constitutes evidence of discrimination.").

56. This is so if the conditions giving rise to the data, the method of data collection, and the alternative hypothesis do not change. In comparing the results of two different experiments or of observational studies (which may lack randomization and controls), one must consider far more than the P-values for each set of results. Within the context of one experimental or observational design, however, lower P-values indicate stronger statistical evidence for the alternative hypothesis. Thus, contrary to what may be inferred from loose statements like those in note 55, *supra*, data that does not rise to some preordained level of significance is still evidence, and it may be fairly good evidence at that. *But see* Meier, Sacks & Zabell, *supra* note 21, at 152 ("If a difference does not attain the 5% level of significance, it does not deserve to be given weight as evidence of a disparity. It is a 'feather.'").

for believing that a proposition is true simply because the evidence for it is "statistically significant" at the .05 level.[57] Thus, instead of dismissing the statistical disparities that did not attain "significance" at the .05 level and relying entirely on other evidence,[58] the trial and appellate courts in *Segar* should have considered the actual magnitudes of the P-values. Statistical evidence need not be dispositive to be helpful in building a prima facie case.

B. Designating the Null Hypothesis

In addition to the difficulty in justifying the choice of a level of statistical significance, there is a further problem. Using a significance level like .05 puts the burden of proof, so to speak, on the proponent of the alternative hypothesis. In most situations, this hypothesis will not be accepted unless there is strong evidence against the null hypothesis.[59] Why should the null hypothesis have this advantage over the alternative hypothesis? A court or jury not fully conversant with statistical terminology could think that experiments or observations that do not uncover any "significant" differences supply decisive evidence that no real difference exists.[60]

C. Misleading Terminology

Another reason for excluding, or at least clearly explaining, testimony that the statistical data are "not significant," "significant," or "highly

57. Meier, *Damned Liars and Expert Witnesses*, 81 J. AM. STATISTICAL ASS'N 269, 270–71 (1986). Fisher's views on the use of significance levels in scientific inference may be worth restating. As indicated in text accompanying note 53, *supra*, he recognized that the choice of the .05 level is arbitrary. He also believed that results said to be significant at any level should not ipso facto be taken as proving the existence of a scientific phenomenon. R. FISHER, THE DESIGN OF EXPERIMENTS 13–14 (9th ed. 1971) ("It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result.").

58. *See supra* note 49.

59. Even the "inexorable zero," which the courts took to be dramatic evidence of discrimination in the days before hypothesis testing in court, may not be sufficient to warrant rejection of the null hypothesis at the .05 level. *E.g.*, Capaci v. Katz & Besthoff, 711 F.2d 647, 654 (5th Cir. 1983). To some extent, however, this depends on what one takes the alternative hypothesis to be. *See* Rubinfeld, *supra* note 34, at 1056–62.

60. In Williams v. Florida, 399 U.S. 78 (1970), the Supreme Court cited empirical research (of dubious quality) on the functioning of twelve-member as opposed to six-member juries. Emphasizing the failure of these limited studies to discern any significant difference between the two types of juries, the Court placed the burden of empirical proof on the wrong party. *See* Lempert, *Uncovering 'Nondiscernible' Differences: Empirical Research and the Jury-Size Cases*, 73 MICH. L. REV. 643 (1975); *cf.* Kaye, *supra* note 16 (pointing to a similar error in Hazelwood School Dist. v. United States, 433 U.S. 299 (1977)).

significant"[61] is that in the context of hypothesis testing these terms lack their ordinary meaning. The magnitude of an observed disparity does influence the P-value. But a P-value is not a direct measure of the magnitude of an observed disparity, and it provides no necessary indication of the importance of an observed difference. With small samples, large differences can be "insignificant." Apparently, this happened with some of plaintiffs' statistics in *Segar*.[62] Conversely, with large samples, picayune differences can be "significant."[63] For example, statistical analysis might show that science majors receive "significantly" better grades in law school than liberal arts students, but if the difference were only a hundredth of a point on a 4.0 scale, no one should care very much about this "significant" difference. *Segar,* which produced one of the best opinions on proof of salary disparities by regression analysis, speaks of "evidence of significant discrimination"[64] when what is meant, presumably, is "significant evidence of discrimination."[65] The ease with which the language of significance testing can be misunderstood is one more reason to steer clear of this terminology.[66]

Difficulty with the language of significance testing is especially telling in jury trials. Most judges, upon study or reflection, can appreciate the distinction between statistical signficance and practical importance.[67] However, most untutored jurors probably will not recognize that an expert's testimony that certain statistical proof is "highly significant" may not mean that a substantial effect has been observed. To be sure, the opposing party can elicit the distinction by cross-examination or through its own experts, but this generally is an imperfect and costly palliative. The result of a significance test or an unadorned statement of the P-value is not itself evidence. Each is merely expert testimony admitted to assist the fact finder

61. *See, e.g.*, Geller v. Markham, 635 F.2d 1027, 1032 (2d Cir. 1980) (expert characterized proportion as "very significant" statistically, about "600 times the level generally required for statistical significance").

62. *See supra* note 49.

63. Rubinfeld, *supra* note 34, at 1067–68.

64. *Segar*, 738 F.2d at 1263.

65. The estimated values of the parameter associated with the variable for race tended to be on the order of $1,000, as in the one regression described in Part II. If a coefficient of this magnitude is large enough to be considered a gross disparity—and it probably is—then it is correct to refer to it as evidencing "significant discrimination." This may be precisely what the court of appeals had in mind when it used the phrase. Given the ambiguity of the word "significant," however, it is impossible to know whether the court characterized the disparity as "significant" because the observed coefficient was large, because its P-value was under .05, or both.

66. This aspect of significance testing has not escaped the attention of social scientists. *See, e.g.*, Skipper, Guenther & Nass, *The Sacredness of .05: A Note Concerning the Uses of Statistical Levels of Significance in Social Science*, 2 AM. SOCIOLOGIST 16, 17 (1967).

67. *See, e.g.*, Bilingual Bicultural Coalition on Mass Media, Inc. v. FCC, 595 F.2d 621 (D.C. Cir. 1978) (Robinson, J., dissenting).

in evaluating the statistical data. A pronouncement that the evidence is "statistically significant" adds nothing of substance to a precise statement of the P-value. The dangers of confusion, misleading the jury, and undue time-consumption, which can make even relevant evidence inadmissible under Rule 403,[68] outweigh the negligible probative value of testimony about "significance" or "hypothesis tests."[69]

Like "significance," "confidence" is a technical term with a meaning that is not what most people think—even those with introductory training in statistics. Despite criticism,[70] some statisticians continue to speak of the "confidence" that a decisionmaker can have in the result of a hypothesis test. This confidence is simply one minus the significance level. Thus, statements like the following appear: "when led to a rejection of the null hypothesis at a level of significance of .05, a court can be at least 95% confident that a disparity of treatment of the relevant groups exists."[71] It should come as no surprise that the judges, who are offered such advice, accept and propagate these characterizations.[72]

Unfortunately, significance probabilities do not translate so freely into expressions of subjective certitude. The probability that the alternative hypothesis is true is *not* generally equal to one minus the significance probability.[73] As the California Supreme Court discerned in the slightly

---

68. *See* MCCORMICK, *supra* note 4, § 185.

69. I am assuming that the expert witness must present the P-value and explain the idea behind this number rather than merely assert that it is "significant," so that the marginal probative value of the expert's imprimatur of significance is *de minimus*. A clear statement of the P-value seems essential if (a) the expert is to follow good statistical practice, and (b) the factfinder is to have any chance of comprehending what the expert means when he or she characterizes the statistical evidence as "significant." *See* Lempert, *Statistics in the Courtroom: Building on Rubinfeld*, 85 COLUM. L. REV. 1098, 1101–02 (1985). For these reasons, conclusory testimony about "significant" results (testimony that does give a reasonably explained P-value) should be inadmissible.

70. *E.g.*, Chandler, *The Statistical Concepts of Confidence and Significance*, 54 PSYCHOLOGY BULL. 429 (1957).

71. Braun, *supra* note 3. Comparable misstatements may be found in D. BARNES, *supra* note 4, at 162; Barnes, *A Common Sense Approach to Understanding Statistical Evidence*, 21 SAN DIEGO L. REV. 809, 831 (1984); Cohen, *Confidence in Probability: Burdens of Persuasion in a World of Imperfect Knowledge*, 60 N.Y.U. L. REV. 385, 401 (1985).

72. In Craik v. Minnesota State Univ. Bd., 731 F.2d 465, 476 n.13 (8th Cir. 1984), the majority of the panel wrote that "[a] finding that a disparity is statistically significant at the 0.095 or 0.01 level means that there is a 5 per cent. or 1 per cent. probablility, respectively, that the disparity is due to chance." Judge Swygert, whose dissenting opinion included an extended discussion of regression methodology, replete with graphs and tables, stated that since "each coefficient was statistically significant at the 1% level . . . we can be 99% confident that each was different from zero." *Id.* at 510. For other examples of this fallicy, see Vasquez v. Hillery, 106 S. Ct. 617, 621 (1986); Rivera v. City of Wichita Falls, 665 F.2d 531, 545 n.22 (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).

73. For a recent reminder of this point, see Fisher, *Statisticians, Econometricians, and Adversary Proceedings*, 81 J. AM. STATISTICAL ASS'N 277, 280 (1986); *cf.* DeGroot, *Doing What Comes Naturally: Interpreting a Tail Area Probability As a Posterior Probability or a Likelihood Ratio*, 68 J. AM.

bizarre case of *People v. Collins*,[74] if the probability that a randomly selected person will fit an eyewitness's accurate description of a robber is as small as 1/12,000,000, it does not follow that the probability that this person is the robber exceeds 0.99999. In a sufficiently large population, several people may fit the same description.[75]

There is, of course, a reason for using the word "confidence" to denote the complement of the significance level. It relates to the notion of a "confidence interval." A "confidence interval" is an estimate of a parameter stated as a range of values that the unknown parameter might have. Such an interval estimate has two components—the interval within which the parameter is reported to lie, and the "confidence coefficient." This confidence coefficent helps determine the width of this interval and it equals one minus the significance level for a particular hypothesis test.

For example, suppose that a simple random sample selected in a public opinion poll commissioned to support a change of venue motion shows that sixty-five percent of the people questioned have the impression that the defendant is guilty. Suppose further that in view of the sample size, this finding leads to an estimate, with a ninety-five percent confidence coefficient, that between sixty and seventy percent of the population share this impression. To test at the .05 level whether the null hypothesis that the proportion of the population leaning toward guilt is any particular number (say fifty percent), we need only ask if that number lies within the interval estimate. If it does not (as is the case for fifty percent), then the sample proportion warrants rejection at the .05 level of the claim that the population proportion is the hypothesized number. But, contrary to what some courts might think,[76] the confidence coefficient of ninety-five percent for this estimate does not mean that it is ninety-five percent probable that the population proportion is between sixty and seventy percent. The ninety-five percent "confidence" pertains only to the statistical procedure that generates an interval estimate. The confidence coefficient of ninety-five percent means that if a great many simple random samples had been taken and a

STATISTICAL ASS'N 966 (1973) (describing conditions under which the common fallacy turns out to be correct.)

74. 68 Cal. 2d 319, 438 P.2d 33, 66 Cal. Rptr. 497 (1968).

75. *See, e.g., Collins*, 66 Cal. Rptr. 497 (1968); Meier, Sacks & Zabell, *supra* note 21, at 149 n.40; Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329 (1971). For more illustrations of the distinction between the P-value level and the probability on which the case should turn, see Kaye, *Statistical Significance and the Burden of Persuasion*, *supra* note 31.

76. *E.g.*, Vuyanich v. Republic Nat'l Bank, 505 F. Supp. 224 (N.D. Tex. 1980). Again, in view of the explanations that appear in law reviews and treatises as well as in court, one can hardly blame the courts for having this impression. *See, e.g.*, D. BARNES, *supra* note 4, at 35; W. LOH, SOCIAL RESEARCH IN THE JUDICIAL PROCESS: CASES, READINGS AND TEXT 410 (1984); Cohen, *Confidence in Probability: Burdens of Proof in a World of Imperfect Knowledge*, 60 N.Y.U. L. REV. 385 (1985); Sprowls, *The Admissibility of Sample Data into a Court of Law: A Case History*, 54 UCLA L. REV. 222 (1957).

different confidence interval computed for each such sample, about ninety-five percent of these intervals would have included the unknown parameter. From the viewpoint of classical statistics, with its frequency based interpretation of probability, this does not imply that the parameter has a ninety-five percent chance of being in any particular interval, such as the sixty percent to seventy percent one.[77] Because this point is difficult to grasp, testimony about "confidence" flowing from significance tests or about the "confidence coefficient" of interval estimates promises to be more misleading than edifying. If so, such testimony should be excluded from the expert's presentation.[78]

D. Searching for Significance

Repeated applications of significance testing confuse the interpretation of a significance level even more. Research that fails to uncover significance tends not to be published. From the viewpoint of other researchers, this can be troublesome, since it unwittingly may condemn them to repeat the search for an effect that does not exist.[79] From the perspective of an attorney looking for an impressive footnote, this bias is not so bad because if enough studies are conducted, statistical error almost guarantees that some will come out the desired way even if there is no real effect.[80]

---

77. *E.g.*, V. BARNETT, COMPARATIVE STATISTICAL INFERENCE 36–37 (2d ed. 1982); Aickin, *Issues and Methods in Discrimination Statistics*, in STATISTICAL METHODS IN DISCRIMINATION LITIGATION 168 (D. Kaye & M. Aickin eds. 1986).

78. This is not to say that the confidence intervals themselves should be excluded. Quite the contrary, when the confidence interval can be computed it should be displayed, for it has several advantages over a statement of the P-value. First, a confidence interval is more revealing than a P-value and includes all the information that is present in the P-value. Second, a confidence interval does not assign the null hypothesis to one party or the other. Finally, the width of the interval is a graphic measure of the probative value of the statistical evidence. These thoughts are developed further in Part V.

When a confidence interval is used in court, however, it should not be denominated a "confidence" interval because the confidence coefficient does not equal the subjective confidence that one should have in the truth of a relevant proposition. The more neutral phrase "interval estimate" might be used, and the "confidence coefficient" referred to simply as a "frequency coefficient" for that estimate.

79. *E.g.*, Zeisel, *The Significance of Insignificant Differences*, 19 PUB. OPINION Q. 319 (1955). The following parable has been used to illustrate the point:

> There's this desert prison, see, with an old prisoner, resigned to his life, and, a young one just arrived. The young one talks constantly of escape, and, after a few months, he makes a break. He's gone a week, and then he's brought back by the guards. He's half dead, crazy with hunger and thirst. He describes how awful it was to the old prisoner. The endless stretches of sand, no oasis, no signs of life anywhere. The old prisoner listens for a while, then says, "Yep, I know. I tried to escape myself, twenty years ago." The young prisoner says, "You did? Why didn't you tell me, all these months I was planning my escape? Why didn't you let me know it was impossible?" And the old prisoner shrugs, and says, "So who publishes negative results?"

J. HUDSON, A CASE OF NEED (1968), as quoted in Walster & Cleary, *A Proposal for a New Editorial Policy in the Social Sciences*, AM. STATISTICIAN, April 1970, at 16, and in D. MOORE, *supra* note 54, at 293.

80. There are some situations in which the opposite problem arises. *See* A.W.F. EDWARDS,

To illustrate how this can happen, consider the problem of deciding whether a coin is biased. The probability that a fair coin will produce ten heads when tossed vigorously ten times is $(1/2)^{10} = 1/1024$. Observing ten heads for the first ten tosses would therefore be strong evidence that the coin is biased. Since the P-value of $1/1024$ is less than .05, one could say that these observations are statistically significant at the .05 level (and at much smaller levels as well). Nevertheless, if a fair coin is tossed a few thousand times, it is quite likely that at least one string of ten consecutive heads will appear.

This problem can develop, probably in more virulent form, in testimony about the more elaborate statistical models mentioned in Part II. Almost any large data set—even pages from a table of random digits—will contain some unusual pattern[81] that sufficient computer time and ingenuity will discover.[82] Having detected that pattern, the analyst who performs a specific test for it will find statistical significance. But like a string of ten heads in thousands of coin tosses, which has a P-value of just under .001 when viewed in isolation, this result proves nothing.

Once one becomes aware of it, the problem of interpreting multiple P-values or significance tests obtained from the same set of data seems ubiquitous. In *Certified Color Manufacturers Association v. Mathews*,[83] for example, manufacturers of food additives disputed the claim that the coloring agent popularly known as red dye number two is carcinogenic. The Food and Drug Administration, in terminating its provisional approval of the substance, relied on a controlled (but poorly executed) two-and-a-half-year experiment in which its scientists randomly assigned rats to four groups, and fed each group a diet having a different concentration of the

---

LIKELIHOOD: AN ACCOUNT OF THE STATISTICAL CONCEPT OF LIKELIHOOD AND ITS APPLICATION TO SCIENTIFIC INFERENCE 180 ("[s]equential rather than concentrated assaults on the null hypothesis are practically powerless in difficult cases; it is like trying to sink a battleship by firing lead shot at it for a long time.").

81. D. MOORE, *supra* note 54, at 294. Thus, it has been reported that murderers generally have long, narrow noses and slit-like mouths, and that suicides tend to occur when atmospheric ozone levels are falling. Curry, *The Relationship of Weather Conditions, Facial Characteristics and Crime*, 39 J. CRIM. L. & CRIMINOLOGY 253, 259 (1948). A more recent survey purported to show a remarkable correlation between using an IBM personal computer and craving pepperoni pizza. 1 PC MAG. 59 (Apr. 1983).

82. *See, e.g.*, Diaconis, *Theories of Data Analysis: From Magical Thinking Through Classical Statistics*, in EXPLORING DATA TABLES, TRENDS, AND SHAPES 8–9 (D. Hoaglin, F. Mosteller & J. Tukey eds. 1985). This problem arises frequently in multiple regression with many variables. *See* Denton, *Data Mining As an Industry*, 67 REV. ECON. & STATISTICS 124 (1985); Freedman, *A Note on Screening Regression Equations*, 37 AM. STATISTICIAN 152 (1983). Here, the intuition of many courts—which suggests that the more variables that are included in the model, the better—leads them astray. *See, e.g.*, McCleskey v. Zant, 580 F. Supp. 338 (N.D. Ga. 1984), *rev'd on other grounds sub nom*, McCleskey v. Kemp, 753 F.2d 877 (11th Cir. 1985), *cert. granted in part*, 106 S. Ct. 331 (1986).

83. 543 F.2d 284 (D.C. Cir. 1976).

additive. Some rats died before the study ended; the rest were killed and examined at the close of the experiment. A biostatistician analyzing the results reported that "it appears that feeding FD&C Red No. 2 at a high dosage results in a statistically significant increase in a variety of malignant neoplasms among aged Osborne-Mandel female rats."[84] One senses that a series of hypothesis tests were performed, but only those involving certain types of tumors and certain types of rats in the control and treatment groups showed statistically significant associations. To sustain the agency's action, this may have been evidence enough, but the multiple testing (not to mention the logical hiatus between a P-value and subjective confidence in the alternative hypothesis) implies that it would be a mistake to think that this experiment established that there is a probability of .95 or more that high doses of red dye number two cause cancer in rats.

Multiple testing was present in *Moultrie v. Martin,*[85] the very case in which the Fourth Circuit imposed its requirement of hypothesis testing and, acting as its own statistician, purported to show that petitioner's evidence of discriminatory grand jury selection was not statistically significant. The *Moultrie* variation given in Part II presented only part of the data.[86] On appeal from the denial of a post-conviction petition for habeas corpus, the Fourth Circuit Court of Appeals tabulated statistics on the representation of blacks on grand juries over a seven-year period. Using the "standard deviation analysis" mentioned in Part I,[87] the court reported the following values for the t-statistic (the number of standard deviations from the mean of a hypothetical distribution associated with the null hypothesis): -3.4, -.9, -.9, .1, .1, -1.4, -1.8. Despite its rhapsodic discussion of hypothesis testing,[88] the court did not perform a formal test to see whether this sequence of outcomes was signficant. Instead, it eyeballed the numbers, gave little weight to the earliest year, which had the largest disparity, and concluded that the serial t-statistics did not show discrimination. Had the court thrown out the first year entirely and performed a hypothesis test on the remainder of the data, it would have had to report that, given the

84. *Mathews,* 543 F.2d at 290.

85. 690 F.2d 1078 (4th Cir. 1982).

86. In addition, the actual case did not involve any "secret" or "alternative" list of registered voters. The alternative hypothesis is therefore more complex than the one used in Part II.

87. For descriptions of the mechanics of the so-called "standard deviation analysis" that seems to have captured the imagination of the courts, see, *e.g.,* authorities cited in Kaye, *supra* note 16, at 837 n.21. The cited authorities indicate some of the limitations of the "standard deviation analysis."

88. *See supra* Part I.

probability model it was using, the statistics *were* statistically significant at the .05 level.[89]

As this discussion indicates, there are some statistical methods for coping with multiple P-values that permit meaningful hypothesis testing in certain cases.[90] But no truly general solution is known,[91] and the existing methods would be of little avail in the typical case where a regression analyst has run through a variety of models to arrive at the one the analyst considers the most satisfactory. In these situations, attorneys and courts should not be overly impressed with claims that the observed coefficient or other quantity of interest is "significant." Instead, they should be asking how the analyst developed and refined the proposed model.

E. Assessing the Model

In evaluating the usefulness of hypothesis testing, it is important to understand that what is being tested is generally limited to a statement about a parameter within the context of a probabability model. For instance, in the modified version of *Moultrie v. Martin* introduced in Part II, the null hypothesis was H_0: $\theta = .38$. This is a claim about the parameter θ, the chance of selecting a black for the grand jury on each draw from the voter list. This parameter is embedded in a model that postulates that every draw is independent, and that the probability of drawing a black grand juror is fixed. The hypothesis test is designed to let us conclude something about θ—it tells us nothing about the model's validity. The alternative hypothesis is not that the model is wrong. It is that the model is right—selection was random with a fixed probability—but that the alternative list was used, so that the model's parameter, θ, is .15 rather than .38.

Yet, the model almost surely is wrong.[92] Even if the jury commissioners

89. Kaye, *supra* note 5. Even so, the court, pursuing the logic of hypothesis testing, might well have concluded that petitioner was not entitled to prevail. Petitioner did not provide evidence of the proportion of blacks who were registered voters in any year except 1977, the year that he was indicted and tried. The court's null hypothesis was that this population parameter was the same in the preceding six years. If the proportion of registered blacks in the South Carolina county was on the rise from 1971 to 1977, the resulting P-value (which is not far below .05) is understated.

90. *See, e.g.*, R. MILLER, SIMULTANEOUS STATISTICAL INFERENCE (2d ed. 1981); Follett & Welch, *Testing for Discrimination in Employment Practices*, 46 LAW & CONTEMP. PROBS. 170 (Autumn 1983); Gastwirth, *Statistical Methods for Analyzing Claims of Employment Discrimination*, 38 INDUS. & LAB. REL. REV. 75 (1985); Kaye, *supra* note 5; Petrondas & Gabriel, *Multiple Comparisons by Rerandomization Tests*, 78 J. AM. STATISTICAL ASS'N 949 (1983).

91. *See, e.g.*, Aickin, *supra* note 77.

92. The model posits what is technically known as a Bernoulli process, and it gives rise to a binomial distribution for the number of blacks selected as grand jurors. A more exact model would recognize that the probability of drawing a black name changes as the number of voters not yet picked for jury service changes with each selection. The distribution generated by this more realistic model would be hypergeometric. Oddly, the courts seem to prefer the Bernoulli model, and have devised their

in the actual case were discriminating, the notion that they were doing so through an "alternative list" is slightly absurd.[93] Identifying the null hypothesis with "no discrimination" and the alternative hypothesis with "discrimination," as some courts are wont to do,[94] is valid only if the alternative hypothesis is part of a probability model that resembles the process of discrimination.[95]

Similar remarks apply to more complex statistical models. The analyst postulates a model with a certain mathematical structure. The analyst then "tunes" the model to fit the data. Finding a decent fit tends to confirm the choice of model. Hypothesis tests, however, usually concern the parameters of the model without addressing the reasonableness of the model itself.[96] Furthermore, when more than one model is advanced, such as when there is an argument about the number of intercorrelated variables that should be put into a multiple regression equation,[97] or when there is a dispute over the value of doing cohort analysis instead of regression analysis,[98] there is no simple or single mathematical test for deciding which

---

own rules for handling small samples. *E.g.*, EEOC v. Federal Reserve Bank, 698 F.2d 633, 650 (4th Cir. 1983). Although the technical objection to the Bernoulli model can be important in employment discrimination cases, see Kaye, *supra* note 21, in the typical jury selection case, the specific binomial and hypergeometric distributions usually are almost identical. *See* Kaye, *supra* note 5.

93. On the other hand, it might be that the commissioners would always summon a white when one was randomly picked, and would summon every other black whose name randomly appeared. The appropriate statistical model for this process differs from the one presented for the case of the "alternative list."

94. *E.g.*, EEOC v. American Nat'l Bank, 652 F.2d 1176, 1192–93 (4th Cir. 1981) ("chance" versus "the only other hypothesis—discrimination").

95. *Cf.* Rubinfeld, *supra* note 34, at 1056–62 (importance of specifying alternative hypothesis correctly).

96. V. BARNETT, *supra* note 77, at 31; Meier, Sacks & Zabell, *supra* note 21, at 152–53. An appendix in Landes & Posner, *Joint and Multiple Tortfeasors: An Economic Analysis*, 9 J. LEGAL STUD. 517, 552 (1980), illustrates the point. The authors use a multiple regression model to show that statutes that permit contribution among joint tortfeasors (which the authors regard as less economically efficient than the common law rule of no contribution) are more likely to be found in states with public policies that generally sacrifice efficiency. Examining the t-statistics for the regression coefficients, Landes and Posner conclude that their statistical analysis "indicates a positive and significant relationship between the government-expenditures variable [used to measure a state's proclivity for inefficient policies] and the probability that a state allows contribution." Although they report that the fitted regression equation has an R-square of only .09, they never test the hypothesis that there simply is no regression relationship of the type they presuppose. Yet the ordinary least square regression model, which is what they appear to have used, is inferior to a logistic model when the dependent variable is binary. Campbell, *Regression Analysis in Title VII Cases*, 36 STAN. L. REV. 1299 (1984), calls attention to this type of problem, but the emphasis on R-square as a solution is misguided.

97. *E.g.*, Valentino v. United States Postal Serv., 511 F. Supp. 917 (D.D.C. 1981), *aff'd*, 674 F.2d 56 (D.C. Cir. 1982); Presseisen v. Swarthmore College, 442 F. Supp. 593 (E.D. Pa. 1977).

98. Segar v. Smith, 738 F.2d 1249, 1263, 1285–86 (D.C. Cir. 1984); Trout v. Hidalgo, 517 F. Supp. 873 (D.D.C. 1981), *aff'd sub nom.*, Trout v. Lehman, 702 F.2d 1094 (D.C. Cir. 1983), *vacated*, 465 U.S. 1056 (1984); Valentino v. United States Postal Serv., 511 F. Supp. 917 (D.D.C. 1981), *aff'd*, 674 F.2d 56 (D.C. Cir. 1982).

model is superior.

This is not to say that standards for evaluating the appropriateness of a given model do not exist. They do, and there are even some hypothesis tests that can be helpful.[99] Knowledgeable statisticians may well reach the same conclusions in a particular case. But as we move into these matters, we leave the simplicity of a single hypothesis test for a particular parameter far behind. There will be disputes among statisticians about "reasonableness" or "appropriateness," which may begin to sound suspiciously like the courtroom exchanges among psychiatrists and other experts from the "softer" sciences.[100] There may be only one right answer, but no known mathematical algorithm will produce it.[101]

F. Contemplating the Alternatives

In discussing significance testing, I have traveled a path obscured by specialized vocabulary and concepts. It may be helpful to summarize the route. First, I have argued that the choice of the significance level—the point at which we will reject the null hypothesis—is outside the scope of simply applying a given test to the data to see whether the numerical evidence is "statistically significant." The mechanical quality of the hypothesis test itself may seem to ensure objectivity, but unless the selection of the significance level is also objective and sensible, this seeming objectivity is illusory. Second, I have suggested that designating a particular hypothesis to be the "null hypothesis" for testing at a demanding significance level gives an advantage to the party whose position is consistent with the alternative hypothesis—an advantage that may interfere with the law's allocation of the burden of persuasion. Third, I have argued that terms like "significant" and "confident" are misleading, since they pertain merely to the reproducibility of results. In view of these problems, I have suggested that these terms be banished from courtroom discourse. The trier of fact is better served by a clear statement and explanation of the P-value or an interval estimate, than by a statistician's characterization of a particular P-value as "significant" or "not signficant." Beyond this, I have warned against being taken in by significance tests or P-values that are obtained

99. *See, e.g.*, D. BELSLEY, E. KUH & R. WELSCH, REGRESSION DIAGNOSTICS: IDENTIFYING INFLUENTIAL DATA AND SOURCES OF COLLINEARITY (1980); S. WEISBERG, APPLIED LINEAR REGRESSION (2d ed. 1985).

100. Courts that are sensitive to these matters find little solace in the seeming objectivity of hypothesis testing. *See, e.g.*, Presseisen v. Swarthmore College, 442 F. Supp. 593, 619 (E.D. Pa. 1977) ("It seems to the Court that each side has done a superior job in challenging the other's regression analysis, but only a mediocre job in supporting their own. In essence, they have destroyed each other and the Court is, in effect, left with nothing.").

101. See Fisher, *supra* note 73, at 279, for suggested procedures for building models that can be defended in court.

after a clever or crude search for something significant in the data. Finally, I have pointed out that the typical hypothesis test or P-value looks at the value of a parameter rather than at a model's reasonableness.

This last point may seem obvious. Statistics—significant or otherwise—derived from an inappropriate model give useless answers. As trite as this observation may be, it bears on whether it is desirable to drag the jargon and mechanics of full blown hypothesis testing into legal disputes. This mode of discourse can obscure the fact that there are always other alternatives besides the one the statistician identifies as H_1 in formulating the test.[102]

Courts are quite capable of appreciating the limited context of the hypothesis test. In *Mapes Casino, Inc. v. Maryland Casualty Co.*,[103] for example, the court recognized the importance of the "extrinsic" alternatives that the proponent of the statistical evidence failed to enumerate. In this case, the plaintiff sought to quantify the amount of its loss due to employee defalcation. The plaintiff casino showed that over an eighteen-month period, the win percentage at its craps tables was 6.37 percent as compared to an expected value (under the null hypothesis) of twenty percent. Although no P-value was computed, the probability of a discrepancy of at least this size would be very small under the null hypothesis, making it reasonable to reject that hypothesis. But what does this prove? The court reasoned that the statistics were probative of the fact that something was wrong at the craps tables, but it held that this demonstration could be used only to corroborate other evidence as to the quantum of damages. The court pointed to other extrinsic hypotheses—such Runyonesque activities as "skimming," "scamming," and "crossroading"—that might have accounted for the losses.[104]

Likewise, in *Moultrie,* it is not hard to see that rejection of the null hypothesis (that each registered black has a thirty-eight percent chance of appearing on a grand jury) does not necessarily imply that the jury commissioners discriminated. Perhaps the commissioners drew names randomly from the voting list but then properly excluded a higher proportion of black voters than white voters because a higher proportion of black voters were illiterate, felons, or otherwise unqualified to serve. Perhaps the commissioners summoned blacks at the rate of thirty-eight percent, but relatively more blacks than whites failed to respond to the summonses.[105]

102. *See, e.g.,* Meier & Zabell, *supra* note 28 (enucleating such hypotheses in a forgery case). Outside the legal realm there are many intriguing examples of the tendency to think that an outrageously small P-value is definitive proof of an alternative hypothesis, even though there are extrinsic alternative hypotheses that are no less plausible than the alternative used in arriving at the P-value. *See, e.g.,* C. HANSEL, ESP: A SCIENTIFIC EVALUATION (1966).

103. 290 F. Supp. 186 (D. Nev. 1968).

104. *Mapes Casino,* 290 F. Supp. at 193.

105. Since these hypotheses are not part of the probability model, the hypothesis test cannot reject

Identifying such extrinsic hypotheses is not a technical procedure. It is the product of practical judgment, combined with an understanding of how the jury selection process should work. In such cases, the legal community is not likely to surrender to the siren song of a successful significance test.

For these reasons, I would not go so far as to say that the problems of extrinsic alternatives and searching for significance are decisive arguments against using P-values or explicit hypothesis tests. Rather, I present these problems to reinforce the salutory tendency of the more perceptive courts to recognize the variety of possible alternatives to a null hypothesis—not just the "alternative hypothesis" pertaining to the parameter value.[106]

IV. IMPROVED HYPOTHESIS TESTING

The limitations on hypothesis testing[107] surveyed in Part III should make it plain that when statistical evidence is relevant to the resolution of a disputed factual question in court, the procedure is no panacea. In making this point, I may have been preaching to the converted. It is one thing to say, as some courts have, that hypothesis tests are an "objective" and conventional procedure for statistical inference. It is another to believe that they are all one needs to assess statistical arguments. It is not so clear that any courts have embraced the latter view. Even the Fourth Circuit, while continuing to insist that "a finding of legally significant variations based on statistical evidence may not be made in the absence of a finding of

---

or accept them. Nonetheless, it may be that the appropriate legal rule should not place the burden of disproving these possibilities on the petitioner. After all, a prima facie case can be rebutted. *See, e.g.*, Kaye, *Statistical Evidence of Discrimination*, 77 J. AM. STATISTICAL ASS'N 773 (1982).

106. Those acquainted with the voluminous and sometimes vociferous literature on significance testing in the sciences will recognize that there is nothing very original in this collection of defects or limitations of hypothesis testing. Because forensic statistics is still in its infancy, however, there is some value in reiterating these criticisms of hypothesis testing. The courts should not be condemned to repeat the mistakes of other disciplines that rely on statistical argument and analysis.

107. Some writers distinguish between "hypothesis testing" and "significance testing." *See* V. BARNETT, *supra* note 77, at 129. They use "hypothesis testing" to denote procedures that involve the explicit statement of two hypotheses and a critical region in which the test statistic leads to rejection of the null hypothesis. This decision-oriented approach is associated with the work of J. Neyman and E.S. Pearson. "Significance testing," in contrast, may denote a procedure that assesses the evidence against a hypothesis, without specifying a rule for reaching a decision about that hypothesis. It is what I have been describing as a simple presentation of the P-value, and it seems closer to Fisher's views on statistical inference in science, see Fisher, *Statistical Methods and Scientific Induction*, *supra* note 53, at 471–72 and more in keeping with the expert witness' role in court. *Cf.* Marshall & Olkin, *A General Approach to Some Screening and Classification Problems*, 30 J. ROYAL STATISTICAL SOC'Y SERIES B 407, 440 (1968) (statement of Professor Kerridge in *Discussion on the Paper by Dr. Marshall and Professor Olkin*) ("It is not primarily the responsibility of a statistician to make decisions for other people—not in general at any rate . . . . It is for somebody else to say what decisions should be made with . . . information. In other words, ideally, it is the statistician's job to inform not to decide.").

"statistical significance,"[108] has conceded that "[t]he adoption of a particular level or test of statistical significance, . . . is arbitrary."[109]

Nevertheless, I have done more than simply advance the proposition that hypothesis tests are not all there is to making intelligent decisions on the basis of statistical evidence. I have contended that, in the context of litigation, the consumers of neatly packaged hypothesis tests are more likely to be misled than enlightened. But this claim needs to be qualified. If the price is right, expert testimony will be available to counteract the sources of error that I have mentioned. Thus, the real question for the law of evidence is whether the costs of educating the triers of fact are worth the benefits that formal hypothesis testing can bring to the factfinding process.

Although I have suggested that this question should be answered in the negative,[110] I treated hypothesis tests at an elementary level. While most court presentations probably do not go beyond this level, if hypothesis testing is to be given a fair trial, we should consider its full potential, and not merely an early record that includes unsophisticated or thoughtless applications of the technique. This section considers an addition to hypothesis testing that many statisticians consider superior to the simplified approach outlined in Part II. I conclude, however, that this addition is not adequate to keep formal hypothesis testing viable for forensic use.

The improvement involves attending to the "power" of the test. Remember that the hypothesis test in *Moultrie* led us to accept the null hypothesis when three out of eighteen grand jurors were black. This outcome does not mean that the commissioners used the "null list." It merely reflects the fact that the test has little power to discriminate between the null and the alternative hypothesis. The formal and quantitative method of expressing this characteristic of the test is known as the "power function."[111]

---

108. EEOC v. Federal Reserve Bank, 698 F.2d 633, 648 (4th Cir. 1983).

109. *Id.* at 647 (quoting Smith & Abram, *Quantitative Analysis and Proof of Employment Discrimination*, 1981 U. ILL. L. REV. 33, 43). This language contrasts with the same court's description of hypothesis testing only a few months earlier. *See supra* text accompanying note 20. After *Federal Reserve Bank*, the rule in the Fourth Circuit seems to be that hypothesis testing is a prerequisite to finding discrimination from stastistical evidence, but that the signficance level need not be set at .05 as long as it is "acceptable" on the basis of as yet unstated criteria. On balance, the opinion suggests that the court is moving toward the position that small P-values and substantial disparities are required for there to be statistical proof of discrimination. *But see* Bazemore v. Friday, 751 F.2d 662, 673 (4th Cir. 1984) (misreading *Hazelwood* as establishing "the rule" that "more than two or three standard deviations would be required to undercut the presumption that employment decisions were being made without respect to race."). The insistence of the D.C. Circuit in *Segar* that "significant" disparities are essential, combined with the reluctance of that court to adopt explicitly a specific threshold for determining "significance," suggests that the rule in the D.C. Circuit is similar.

110. *See supra* Part III.

111. The "operating characteristic function," which is mathematically identical to 1 minus the

To see whether testimony on this point might be useful in court, let us reconsider the analysis of the underrepresentation of blacks on the grand jury that indicted Moultrie. As in Part II, we take as the null hypothesis H_0: $\theta = .38$. To formulate the alternative hypothesis, we no longer assume the existence of a single "alternative" list. Following the approach of the Fourth Circuit Court of Appeals (even though the model implicit in this approach is implausible[112]), we assume that the commissioners might have used any one of a vast number of alternative lists. That is, we take the alternative hypothesis to be that θ is something other than .38, though we cannot say how far the true value is from .38. In symbols, we write H_1: $\theta \neq .38$. Given that there were only eighteen grand jurors selected, that we are considering this two-sided alternative hypothesis,[113] and that we want a significance level of .05, the only outcomes that would lead to rejection of the null hypothesis are fewer than three, or more than eleven, black jurors. Intermediate values will not count as "statistically significant" evidence against the hypothesis of random selection from the proper list.

We now ask the following question: For all of the possible values of the parameter θ that represent the chance of selecting a black on each grand juror draw, what is the probability that application of this test will cause us to reject the null hypothesis? This probability, which varies as θ assumes different values, constitutes the power function of the hypothesis test. We

power function, also is used. *See, e.g.*, J. MELSA & D. COHN, DECISION AND ESTIMATION THEORY 32–38 (1978); NATIONAL RESEARCH COUNCIL; COMMITTEE ON EVALUATION OF SOUND SPECTROGRAMS, ON THE THEORY AND PRACTICE OF VOICE IDENTIFICATION 27–30 (1979). This curve represents the risk of failing to recognize the alternative hypothesis as correct when in fact it is correct for each possible value of the unknown paramter.

112. *See supra* note 87.

113. We might have said that the alternative hypothesis is that the commissioners used a list in which blacks were underrepresented to some unknown degree, *i.e*, that $\theta < .38$. This seems more reasonable than thinking that instead of drawing from the correct list, the commissioners drew from one that contained too few whites. Yet the *Moultrie* court, like many others, unthinkingly used a two-sided test. To the extent that the choice between a one-sided and a two-sided alternative hypothesis is often debatable, the use of hypothesis testing may not be quite as objective as it first appears to be. This difficulty arose in EEOC v. Federal Reserve Bank, 698 F.2d 633 (4th Cir. 1983). In this case, the court of appeals seems to say that one-tailed tests are not appropriate, because some statisticians describe them as "data mining." *Id.* at 655. What the textbook cited for this proposition actually says is that deciding to use a one-tailed test after running a two-tailed test is a form of "data snooping," which is "a perfectly reasonable thing to do" if certain precautions are observed. D. FREEDMAN, R. PISANI & R. PURVES, STATISTICS 494 (1978). As these authors point out, it is only "the arbitrary [significance levels] at 5% and 1% which make the distinction between two-tailed and one-tailed tests loom so large." *Id.* at 496. If one looks at the P-value as indicating one aspect of the strength of the statistical evidence, rather than as a number that must exceed some preordained value to warrant some action, "it doesn't matter very much whether an investigator makes a one-tailed or a two-tailed z-test, as long as he tells you which it was." *Id. See also* Goldstein, *Two Types of Statistical Error in Employment Discrimination Cases*, 26 JURIMETRICS J. 32 (1985) (defending one-tailed testing as having greater power than two-tailed testing).

already know that if the null hypothesis, which asserts that the true value of θ is .38, is correct, the probability of mistakenly rejecting H_0 in favor of H_1 is about .05. This is what it means to insist on a significance level of .05. To put it yet another way, if we somehow could apply this test over and over with the model in the *Moultrie* case, we would reject the null hypothesis improperly in no more than one out of every twenty such cases.[114] In short, we know that the test is very good at accepting the null hypothesis when that hypothesis is true. The power function takes us one step further. It indicates how sensitive the test is to rejection of the null hypothesis when that hypothesis is false.

Computing the values of the power function for the test used in *Moultrie* is more tedious than difficult. The results are displayed in Figure 1. As one would expect, the test has little chance of rejecting the null hypothesis when the alternative list is only slightly different than the proper list. But what should give us pause is that the test does not have a better than even chance of correctly detecting the use of an alternative list—unless the list is so grossly biased ($\theta < .15$) that a black's chance of appearing on a grand jury is diluted by some sixty percent.[115]

A court that could recognize this power function and understand its meaning would realize that the failure to find "significance" does not "undercut" or "weaken" the alternative hypothesis.[116] It simply reflects the inability of the test to recognize that the alternative hypothesis is correct when in fact it is correct.[117]

Perhaps presentations along these lines might be useful in some cases.[118]

---

114. Actually, the test is even more sensitive to a false rejection. Rejecting H_0 when the number of blacks is 0–2 or 12–18 amounts to adoption of a significance level of .03. If we were to expand the critical region to include 3 blacks, however, we would be using a level of .06. Since there is nothing in between, speaking of the .05 level in this case is misleading. Anything significant at the .05 level is also significant at the .03 level. We are therefore demanding more than the .05 figure suggests.

115. Sixty percent is the relative difference between the proportion of blacks on the voting list and the proportion on the grand jury. As explained in Kaye, *supra* note 5, it is not the best measure of the degree of underrepresentation, but it is preferable to other measures that the courts have used.

116. The *Moultrie* court is not guilty of this misinterpretation. For an example of such a characterization of data that are not quite statistically significant at an arbitrarily selected significance level, see Hazelwood School Dist. v. United States, 433 U.S. 299, 311 & 311 n.17 (1977).

117. *Cf. supra* note 80 and accompanying text.

118. Henkel & McKeown, *Unlawful Discrimination and Statistical Proof: An Analysis*, 22 JURIMETRICS J. 34 (1981), pursue such an analysis. For data from two discrimination cases, they compute the risk of a miss in testing for significance at the .05 level given particular, hypothetical values of the unknown parameter. In other words, they give, in numerical form, certain points on the operating characteristic curve. They conclude that using a pre-established level of .05 unfairly advantages defendants. *See also* Dawson, *Are Statisticians Being Fair to Employment Discrimination Plaintiffs?*, 21 JURIMETRICS J. 1 (1980).

The matter may be more complex than this. A more general analysis of the properties of hypothesis tests for simple hypotheses, on the basis of data sampled from a normal distribution (which is the

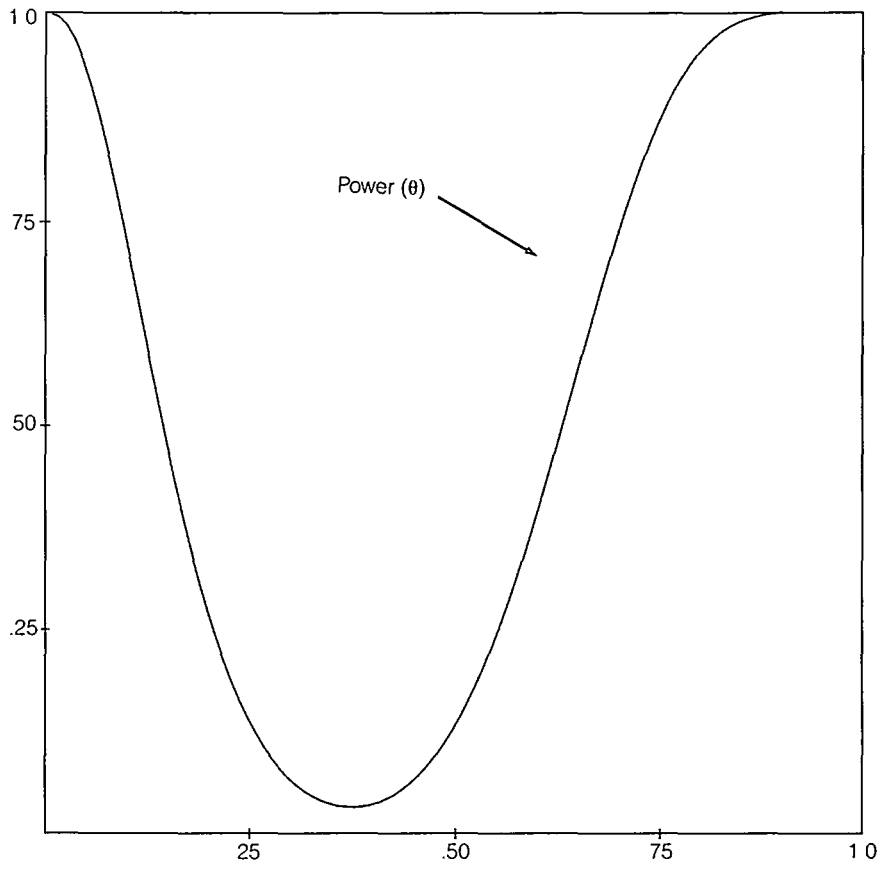

Figure 1. Power Function
for Hypothesis Test in *Moultrie*

For example, pointing out that a test had a power function like that shown in Figure 1 might help a plaintiff counter a defendant's misleading claim that its statistics show quantitatively that the evidence of discrimination is "not significant." But I fear that most of the time talk of "power" would sail smoothly over the heads of the finders of fact. Moreover, such presentations would address only a small portion of the concerns raised in Part III. It may be that courts that permit testimony as to "significance" and "rejection" or "acceptance" of "hypotheses" should also insist on seeing the power functions. Even with this supplement, however, the assistance that the trier of fact might receive from the presentation of hypothesis tests beyond a simple statement of the P-value seems too slight to justify explicit use of the tests in court.

---

approximation that Henkel and McKeown use), reveals that using a fixed significance level of .05 can lead to rejection of H_0 for some samples that actually provide strong evidence (as indicated by the likelihood ratio for these hypotheses) that H_0 is true. M. DEGROOT, *supra* note 2, at 380–81.

Still, the statistical concept of power has other implications, and one commentator has recommended a slightly different application of the power function. Dawson argues that since the civil burden of proof is a preponderance of the evidence, not a "scientific certainty," the "appropriate level of test . . . should be that which equalizes the competing risks . . .[,] the level that balances confidence and power."[119] His proposal, in other words, is to move the significance level to the point where the risks of falsely rejecting the null hypothesis and falsely accepting the null hypothesis are equal, and then to apply the significance test. The power function enters into the formulation of the test procedure, but the function need not be exhibited or explained.

This effort to derive the requisite significance level from the burden of persuasion cannot avoid the criticism that the choice of the significance level is arbitrary and inconsistent with the values that inform the applicable evidentiary standard. Although the preponderance of the evidence standard reflects the principle that the cost of a mistaken verdict for plaintiff is neither greater nor less than the cost of a mistaken verdict for defendant,[120] this standard is concerned with the probability, estimated in light of the evidence in the case, that plaintiff's version of the dispute is correct.[121] Using H_1 to represent plaintiff's version of the facts in dispute, and H_0 to represent defendant's version, we can abbreviate this decisively important probability as $Pr(H_1|Data)$. The preponderance of the evidence standard dictates a decision for plaintiff whenever $Pr(H_1|Data)$ exceeds $Pr(H_0|Data)$, thereby minimizing the probability of a mistaken verdict.

In contrast, the proposal to equate the risk of the two types of errors focuses on two quite different probabilities—$Pr(D_1|H_0)$, the probability of making a wrong decision (accepting H_1) given that the null hypothesis H_0 is true, and $Pr(D_0|H_1)$, the corresponding probability of making a wrong decision (accepting H_0) given that the alternative hypothesis H_1 is true. This procedure sets a threshold that has no necessary connection with $Pr(H_1|Data)$, and it does not keep the probability of an erroneous verdict to a minimum. As a result, setting the significance level according to the error costs generally does not conform to the law's evidentiary standard.[122] The

119. Dawson, *supra* note 118, at 14.

120. *See, e.g.*, Kaye, *The Limits of the Preponderance of the Evidence Standard: Justifiably Naked Statistical Evidence and Multiple Causation*, 1982 AM. B. FOUND. RES. J. 487. This decision-theoretic interpretation of the civil burden of persuasion has proved controversial.

121. Someone who denies the validity or applicability of subjective probabilities to a prescriptive model of the trial process will not accept the claim that the finder of fact can arrive at such a probability. Cohen, *The Role of Evidential Weight in Criminal Proof*, 66 B.U.L. REV. (in press).

122. A more extensive analysis of the relationship between the burden of persuasion and the "equalized" significance level can be found in Kaye, *Hypothesis Testing in the Courtroom*, *supra* note 2.

expert should not be "testing" the statistical evidence at the levels demanded in scientific research or at a level that he thinks the law should require. He should be informing the judge or jury so that these persons can make their own decisions, using the law's standards for evaluating evidence.

V. SOME ALTERNATIVES[123]

A. *The P-Value*

Thus far, I have argued that there is so little to be gained by the trier of fact from being told the result of an hypothesis test, and so much potential for confusion and distraction, that explicit hypothesis testing should not survive a well-developed Rule 403 objection. An expert who can perform an hypothesis test can always do something better. The expert can state the P-value: Properly explained, this number can be of sufficient assistance to the trier of fact to warrant its admission.

Of course, the P-value alone does not establish proof by a preponderance of the evidence, or proof beyond a reasonable doubt.[124] This result is implicit in the distinction, noted in Part IV, between the probabilities to which the preponderance standard applies and those to which a significance test applies. A small P-value (a "significant" or a "highly significant" result, in the terminology I have criticized) does not guarantee "legal significance." It does not always establish that the probability favoring the alternative hypothesis, $Pr(H_1|Data)$, is large.[125] Inversely, a large P-value—a "very insignificant" result—need not imply a small posterior probability $Pr(H_1|Data)$. The data that give rise to a P-value may be too limited for the statistical analysis to be very probative, and may be even more likely to arise under an alternative hypothesis than under the null hypothesis. For instance, in the *Moultrie* example of Part II, the P-value

123. The procedures considered in this section are traditional antidotes to "classical" hypothesis testing. The "likelihood" methods mentioned in Kaye, *supra* notes 2, 16 & 105, also are preferable to explicit hypothesis tests.

124. *See* Kaye, *Statistical Significance and the Burden of Persuasion*, *supra* note 31.

125. Part of an example constructed by the statistician L.J. Savage illustrates this possibility. Savage imagines an inebriated party-goer who says that he can predict the outcome of a coin toss. A coin is tossed ten times, and the party-goer is correct every time. Contrast this with a music expert who says he can distinguish a page of Haydn score from one of Mozart. This individual makes a correct assignment for ten pairs of pages. In each case the P-value is the same, $(\frac{1}{2})^{10} = \frac{1}{1024} < 0.001$, in a one-tail test of significance. Yet, most people probably would accept the musicologist's claim, but dismiss "this drunk's run of luck." L. Savage, *The Subjective Basis of Statistical Practice* (Report, University of Michigan 1961), as described in V. BARNETT, *supra* note 77, at 11–12.

was .051. The analogous probability computed under the alternative hypothesis, that the commissioners picked the grand jurors from the list that was fifteen percent black, is .720. It would be far more probable to find so few blacks on the grand jury under the "not significant" alternative hypothesis than under the null hypothesis.

This last example may appear to suggest that whenever possible, the analyst should report an analog to the P-value, Pr(Extreme Data$|H_1$), along with the P-value. Unfortunately, when the alternative hypothesis involves a broad range of possible values for the parameter in question, this will not be possible. Even here, however, the court can better put statistical proof in proper perspective if it is informed: (a) that the P-value is computed according to a specific probability model; and (b) that, without knowing exactly what alternative model and parameters to use, no expert can tell the court what the probability of finding such data is if, as plaintiff claims, the null hypothesis is false. Conclusory testimony as to "statistical significance" conveys too little in the way information and too much in the way of innuendo.

B. Interval Estimates

Although a clear statement of the P-value is greatly preferable to a blanket assertion of the presence or absence of "statistical significance," there is a procedure that promises to be still more helpful than the P-value approach. Whenever possible, the court should require the expert to give an interval estimate of the parameter in question. As indicated in Part III, the logic of interval estimation is that if one were to repeatedly estimate a parameter on the basis of the many data sets generated by the statistical model, the various estimates would be distributed around the true value of the parameter in a probabilistically well-defined way. One would expect the estimates to fall within a given distance of the unknown, true value a certain percentage of the time. For example, in *Moultrie* the estimated value for θ, the proportion of blacks on the list, was $3/18 = .17$. If more grand juries were drawn randomly from the same list, and if the composition of each such grand jury were used to estimate θ, other estimates would be obtained: some would be higher, and others lower, than .17. For each estimate, if we were to state that the true value for θ lies within a certain range (computed by the same formula for each estimate), and if we wanted to be correct in about half of these interval estimates, then the estimated interval derived from the one grand jury in which three jurors were black would be the observed proportion .17 plus-or-minus .06. Here, the interval estimate is that θ is between .11 and .23, and the process that led to this estimate would give correct results about half the time.

If we wished to use a process that would give correct estimates more frequently, then we would have to be less precise about the value of θ. We would have to say that θ lies within a broader interval about the observed proportion .17. For example, a formula that would give correct estimates in ninety percent of the cases to which it is applied produces an interval estimate of .01 to .32.

One advantage of interval estimation with a variety of confidence coefficients is that it emphasizes that the trier of fact, not the statistician, should decide how accurate the procedure that gives the estimates should be. If a method that would be accurate in half the cases is desired, the statistician states one range of possible values for the proportion of blacks on the list. If a more accurate method is desired, the statistician must give a another range of possible values.

Another advantage of interval estimation is that it gives a range of plausible values for the parameter in question, rather than a single number. If this range is very broad, as in *Moultrie,* then the trier of fact can deduce that the statistical evidence is not very informative. This avoids interference with the law's burden of proof that results from assigning the null hypothesis to one side, and forcing the opposition to disprove the null hypothesis at some preordained significance level that bears no necessary relation to the applicable burden of persuasion. Although the explanation and presentation of interval estimates may be more complicated than a simple statement of the P-value, these estimates convey enough additional information that this price seems worth paying. Courts should move beyond explicit hypothesis testing and P-values, and demand interval estimates whenever possible.[126]

VI. CONCLUSION

This article reflects a particular philosophy about the role of statistical experts in litigation. The underlying premise is that the expert's proper role is not to decide what the statistical evidence proves or disproves. That task, I have supposed, is for the judge or jury. At the same time, statistical evidence cannot be used wisely if it is not understood. The expert can perform an important task by assisting the trier of fact to assess the importance and implications of statistical evidence.

This explanatory function can best be fulfilled by giving the court all the information it needs to evaluate the statistical findings intelligently. Testifying about the result of an hypothesis test does not achieve this ideal. The

126. In some instances, as when nonparametric methods are used, interval estimates cannot be computed. For some cautions about the use of interval estimates, see *supra* note 78.

difficulties with reporting that results are "significant" or "not significant" should require no further reiteration. Statistically significant results may or may not satisfy the applicable legal standard of proof, and trying to construct a test with these standards in mind is not a satisfactory solution to the problems of significance testing.

Presenting the P-value without characterizing the evidence by a significance test is a step in the right direction. Interval estimation, in turn, is an improvement over P-values. With more pieces of the puzzle in hand, the judge and jury stand a better chance of understanding the worth of statistical evidence.

This article is a plea to leave the task of decision to the trier of fact, and not to rely on superficially impressive methods whose seeming objectivity does not withstand analysis. It is a call for using, where suitable, those statistical tools that will aid these decisionmakers in the process of inference. A statistical expert can do no more. He or she should not be allowed to do much less.