Michigan Law Review

Volume 80 | Issue 4

1982

The Numbers Game: Statistical Inference in Discrimination Cases

David H. Kaye
Arizona State University

THE NUMBERS GAME: STATISTICAL INFERENCE IN DISCRIMINATION CASES

*David H. Kaye**

STATISTICAL PROOF OF DISCRIMINATION. By *David Baldus* and *James Cole*. Colorado Springs: Shepard's, Inc. 1980. Pp. xx, 376. $55.

Oliver Wendell Holmes once remarked that "[f]or the rational study of law the black-letter man may be the man of the present, but the man of the future is the man of statistics and the master of economics."[1] To many of his day, this "man of statistics" may have seemed like the perfect attorney for Holmes's quintessential client, "the bad man."[2] The two of them, Holmes might have said, would "stink in the nostrils" of those who would introduce as much fuzziness into the law as they could.[3] Presently, however, there are those who proclaim that Holmes's prophecy has come true.[4] To be sure, the proper role for microeconomic analysis in legal discourse continues to be hotly debated,[5] but few would deny that quantitative methods are becoming increasingly important in litigation. Especially in cases alleging discrimination, judges and commentators have observed that "statistics often tell much, and the Courts listen."[6] Whether this infusion of numerical methods into legal proceedings evokes feelings of approbation or revulsion, it seems clear that at least a rudimentary knowledge of statistical reasoning is essential if attorneys and judges are to function effectively in discrimination litigation.

---

* Professor of Law, Arizona State University. S.B. 1968, Massachusetts Institute of Technology; M.A. 1969, Harvard University; J.D. 1972, Yale Law School. The author wishes to thank Mikel Aickin for thoughtful and insightful comments on a draft of this Review. — Ed.

1. Holmes, *The Path of the Law*, 10 HARV. L. REV. 457, 469 (1897).

2. *Id.* at 459-61.

3. *Id.* at 462.

4. On leaving the deanship at the Stanford Law School, Charles Meyers opined that "without knowing basic economics, lawyers simply will be unable to cope with the last fifth of the 20th century and the next century." *Charles Meyers: Law School Visionary*, CALIFORNIA LAWYER, Oct. 1981, at 47.

5. *E.g.*, Kennedy, *Cost-Benefit Analysis of Entitlement Problems: A Critique*, 33 STAN. L. REV. 387 (1981).

6. B. SCHLEI & P. GROSSMAN, EMPLOYMENT DISCRIMINATION LAW 1162 (1976) (quoting Alabama v. United States, 304 F.2d 583, 586 (5th Cir. 1962)).

This, at any rate, is the thesis propounded by David Baldus, a professor of law at the University of Iowa,[7] and James Cole, a consulting statistician from Pittsburgh.[8] In *Statistical Proof of Discrimination* they argue that quantitative approaches to detecting discrimination have much to offer. Because they recognize that these methods, like other powerful tools, can be dangerous if not handled carefully, they have not written a treatise cataloging and indexing every discrimination case that has discussed statistical proof. Instead, they have crafted a primer for the mathematical neophyte on how statistical techniques ought to be used in ascertaining whether a plaintiff has established a prima facie case[9] of discrimination.[10] The authors employ four hypothetical cases to illustrate various statistical methods, both descriptive and inferential. They weave these paradigmatic cases into the fabric of established decisions. The result is a rich tapestry of quantitative analysis and legal doctrine. Baldus and Cole write at the frontiers of existing case law, and their work will almost certainly exert a powerful influence on the law governing the proof of discrimination.[11]

Of course, none of this means that *Statistical Proof of Discrimination* is beyond the pale of all criticism. The writing is generally careful, but sometimes less than pellucid.[12] At some points, it borders on

7. Professor Baldus, who also served as the Director of the National Science Foundation's Law and Social Sciences Program, has written several important articles on quantitative techniques. *E.g.*, Baldus, Pulaski, Woodworth & Kyle, *Identifying Comparatively Excessive Sentences of Death: A Quantitative Approach*, 33 STAN. L. REV. 1 (1980).

8. Dr. Cole, formerly an assistant professor of statistics at the University of Iowa, also collaborated with Professor Baldus in Baldus & Cole, *A Comparison of the Work of Thorsten Sellin and Isaac Ehrlich on the Deterrent Effect of Capital Punishment*, 85 YALE L.J. 170 (1975).

9. The omission of explicit analysis of the use of statistical evidence in rebutting the prima facie case has occasioned some criticism. *See*, Gruner, Book Review, 49 GEO. WASH. L. REV. 441, 447-48 (1981).

10. Despite the care that Baldus and Cole take in presenting statistical methods in simple terms, the statistically naïve reader should not assume that the book will give him everything he needs to undertake statistical studies, to test whether the assumptions of a statistical model hold in an actual case, or even to hire a statistical consultant wisely. *See* Gerjuoy, Book Review, 66 A.B.A. J. 1100 (1980). For instance, there is no discussion of some common ways in which even a correctly computed correlation coefficient can be misleading, *cf.* A. EDWARDS, AN INTRODUCTION TO LINEAR REGRESSION AND CORRELATION 55-61 (1976) (problems with small samples, combining samples, and restriction of range), or of the power of a classical hypothesis test. *See* note 40 *infra*.

I offer this observation more as a caution to the over-enthusiastic reader than as a criticism of the book. Any nonmathematical treatment of a mathematical subject can convey only a limited understanding (as any reader of *Scientific American* knows). Just as *Statistical Proof of Discrimination* is not an exhaustive legal treatise, so too it is not a comprehensive statistics text. It does, however, contain ample references to such texts.

11. For example, *Statistical Proof* contains the most elaborate analysis yet written for attorneys on how discrepancies between outcomes that would be expected in the absence of discrimination and outcomes that are observed should be quantified. Pp. 144-60.

12. In their introductory assessment of quantitative proof, for instance, Baldus and Cole

the tedious.[13] Although the preponderance of the arguments is convincing, a few are obscure[14] or troublesome.[15] But I do not desire to belabor these occasional flaws or to detail for their own sake the organization or themes of the book. Other reviewers have traveled these roads.[16] In the remainder of this *Review*, I hope to describe more fully the sort of issues with which Baldus and Cole are concerned, to describe how a few of their insights might apply in the context of a specific case, and to consider briefly some approaches that depart from the classical statistical methodology that the authors pursue. I shall begin by describing the Supreme Court's disap-

---

list as an "advantage of statistical proof" the fact that "[i]t can . . . provide a reliable basis for inferring why individuals have been disadvantaged by a selection process," and they refer to the capacity of statistical analysis "to assess causal arguments." Pp. 4-5. They then state that "[t]he primary limitation of quantitative proof . . . is its inability to support an inference about the reasons for a particular decision, such as why a certain individual was hired or fired . . . ." P. 5. Some of the confusion may be engendered by the unexplained use of the word "causal," which has a special meaning in multivariate statistics. *See* Cohn, Book Review, 55 N.Y.U. L. REV. 1295, 1302-10 (1980). I understand Baldus and Cole to be saying two things here: (1) that statistical methods can demonstrate that a selection or allocation process has burdened or benefited one group more than another, but this analysis cannot by itself establish the motivation behind this selection procedure; and (2) that even where the quantitative evidence reveals a difference in the way two groups are treated, a further inference is required to conclude anything about individuals in these groups.

13. Thus, the first chapter on "Discrimination and Models of Proof" (pp. 9-52) enumerates four "theories" (also called "models") of discrimination and another three "models" of proof of the disparate treatment theory or model. Although analytic precision may warrant this proliferation of concepts, one wonders whether the social scientists and statisticians for whom this chapter is written, *see* p. 4, will not find more accessible overviews of the pertinent legal doctrines elsewhere.

14. For example, in arguing that it is usually desirable to measure adverse impact by a simple difference between two numbers (as opposed to a ratio), Baldus and Cole state that "our analysis of the cases suggests that the assumption of the difference measure better approximates the disutility structure underlying the law in more situations than does the assumption of the ratio measure." Moreover, they add, "we suggest that this will continue to be true in the future, although in some respects it is an empirical question depending strictly on the facts of the cases." P. 149. They do not explain how they were able to discern this underlying disutility structure in the unnamed cases they have in mind.

15. *See* note 69 *infra*. Consider also what Baldus and Cole say about the relation of the ratio measure of adverse impact to the utility of money as a function of income:

> The argument for ratios in wage and similar benefit cases rests on the generally accepted premise that a $1,000 deprivation of salary is less important to someone in the $20,000 range than to someone in the $10,000 range. Consequently, a $1,000 deprivation may represent greater harm to someone in the latter group than to someone in the former. Moreover, it may well be that the 5 percent loss of $1,000 to someone in the $20,000 income range represents the same actual harm as would a 5 percent loss of $500 to someone earning $10,000 per year. If this is true, it would also tend to support the use of a ratio measure in these cases, since the percentage loss or relative disparity is simply one form of the ratio measure.

P. 155. The problem is that neither the claim that the utility of money is a logarithmic function of income nor the contention that interpersonal comparisons of utility are meaningful is "generally accepted" by psychologists, economists and others who have investigated and developed the theory of utility. *See, e.g.*, W. BAUMOL, ECONOMIC THEORY AND OPERATIONS ANALYSIS 193-95, 421-32 (4th ed. 1977).

16. *See* Cohn, *supra* note 12; Gerjuoy, *supra* note 10; Gruner, *supra* note 9.

pointing treatment of statistical proof in *Hazelwood School District v. United States*,[17] a well-known employment discrimination case. Second, I shall indicate what the attorney who has mastered *Statistical Proof of Discrimination* might have to say about this unhappy opinion. Finally, I shall examine the statistical evidence in *Hazelwood* with the aid of two theories of statistical inference not seriously considered in *Statistical Proof*.

I. *HAZELWOOD V. UNITED STATES* AND "THE MAN OF STATISTICS"

In *Hazelwood*, the United States brought an action under title VII against a St. Louis County, Missouri, school district. The government alleged that the district was engaging in a "pattern or practice" of discrimination in hiring teachers. To demonstrate the existence and extent of this pattern, the government pointed to, among other things, data showing that although 15.4% of the teachers in the geographical region were black, the comparable proportion among Hazelwood's teaching staff was only 1.4% and 1.8% in 1972-1973 and 1973-1974, respectively.

The district court held that these statistics were "nonprobative" on the curious ground that the percentage of black students in the school district was also trifling.[18] This amounts to saying that a school district can refuse to hire black teachers as long as there are not too many black students around. The court of appeals reversed and directed judgment for the government. It reasoned that the proper comparison was between the proportion of black teachers in the Hazelwood district and the proportion of black teachers in the labor market from which the district drew its teachers. If many black teachers were available for employment, but only few were hired, it would be natural to suspect that the hiring process was biased against blacks. The court of appeals thus held that the figures given above constituted a prima facie (and unrebutted) case of racial discrimination.

The Supreme Court differed with the court of appeals. To be sure, it agreed that the district court's reasoning was "fundamentally misconceived,"[19] which is a polite way to put it, but it questioned whether the relevant labor market included the City of St. Louis. It observed that if the city were excluded from the market, the percent-

17. 433 U.S. 299 (1977).

18. 433 U.S. at 304.

19. 433 U.S. at 308.

age of black teachers would plummet from 15.4 to 5.7, which would put things more in line with the figures for Hazelwood's staff. It therefore vacated the judgment of the court of appeals and remanded the case to the trial court for findings on the scope of the relevant labor market.

To support its intuition that with the City of St. Louis excluded, the statistics might not create a prima facie case, the Court offered a learned footnote seemingly steeped in statistical wisdom:

> [U]nder the statistical methodology . . . involving the calculation of the standard deviation as a measure of predicted fluctuations, the difference between using 15.4% and 5.7% as the areawide figure would be significant. If the 15.4% figure is taken as the basis for comparison, the expected number of Negro teachers hired by Hazelwood in 1972-73 would be 43 (rather than the actual figure of 10) of a total of 282, a difference of more than five standard deviations; the expected number in 1973-74 would be 19 (rather than the actual figure of 5) of a total of 123, a difference of more than three standard deviations. For the two years combined, the difference between the observed number of 15 Negro teachers hired (of a total of 405) would vary from the expected number of 62 by more than six standard deviations. Because a fluctuation of more than two or three standard deviations would undercut the hypothesis that decisions were being made randomly with respect to race . . . each of those statistical calculations would reinforce rather than rebut the Government's other proof. If, however, the 5.7% areawide figure is used, the expected number of Negro teachers hired in 1972-1973 would be roughly 16, less than two standard deviations from the observed number of 10; for 1973-1974, the expected value would be roughly seven, less than one standard deviation from the observed value of 5; and for the two years combined, the expected value of 23 would be less than two standard deviations from the observed total of 15.[20]

To readers not versed in statistics, this footnote must seem formidable indeed. Yet, the essence of what the Court is saying is obvious enough, and an attentive reading of *Statistical Proof of Discrimination* should dispel most of the mystery.[21] Even if an employer makes hiring decisions without regard to race, it is always possible that the proportion of blacks hired will differ slightly from the proportion in the pool of all the applicants. Even large discrepancies are possible, though they are less likely than small ones. This phenomenon of random fluctuation, or sampling error, is familiar enough. After all,

---

20. 433 U.S. at 311 n.17.

21. For Baldus's and Cole's explanation of the *Hazelwood* calculations, see pp. 294-97. *See generally* W. CONNOLLY, JR. & D. PETERSON, USE OF STATISTICS IN EQUAL OPPORTUNITY LITIGATION 74-83 (1979); C. SULLIVAN, M. ZIMMER & R. RICHARDS, FEDERAL STATUTORY LAW OF EMPLOYMENT DISCRIMINATION 78-80 (1980); Braun, *Statistics and the Law: Hypothesis Testing and Its Application to Title VII Cases*, 32 HASTINGS L.J. 59, 72-75 (1980).

if one tosses a balanced coin ten times, there can be no guarantee that exactly five heads will turn up. Similarly, if one blindly draws ten marbles from an urn containing ten black marbles and ninety white ones, there is no guarantee that any particular number so obtained will be black. By chance alone, all the marbles sampled might turn out to be white. Or, several might be black.

The mathematical theory of probability enables us to quantify the chances involved. In the example of the urn, the probability that *no* black marbles will appear in any given sample of ten is a little less than .04. That is, even if the drawings were perfectly fair, the outcome would grossly favor whites (and perhaps appear biased) about four times out of every one hundred.

Similarly, the *Hazelwood* Court is asserting that with the teachers in the City of St. Louis removed from the applicant pool, the seemingly small proportion of black teachers hired is sufficiently close to the proportion in the labor market that the modest discrepancy has a good chance of arising from sampling error rather than from bias in hiring. Indeed, the Court says that the chance is so high that it may "weaken" the government's claim of discrimination.[22]

Equipped with the explanations in *Statistical Proof of Discrimination*, an astute advocate or commentator should find fault with this conclusion on a variety of grounds.[23] At the outset, he might question the entire effort to compare the proportion of black teachers that the Hazelwood district had hired to the proportion of black teachers in the suburban schools as a whole. The comparison ideally should involve only those teachers available for employment with Hazelwood, and ordinarily those who have in fact applied provide the best indication of this pool of potential applicants.[24] The Supreme Court, recognizing that data involving such teachers might be preferable to that used by the government, noted in remanding the case that "[i]t will be open to the District Court . . . to determine whether sufficiently reliable applicant-flow data are available to permit consider-

22. 433 U.S. at 311.

23. Baldus and Cole disapprove of rigid hypothesis testing at the .05 level, although they do not mention this view in their discussion of *Hazelwood*. *See* p. 308. Oddly, most of the commentary of *Hazelwood*, even that which focuses on the Court's use of statistical reasoning, is almost entirely uncritical. *See, e.g.*, F. MORRIS, JR., CURRENT TRENDS IN THE USE (AND MISUSE) OF STATISTICS IN EMPLOYMENT DISCRIMINATION LITIGATION 38-39 (2d ed. 1978); C. SULLIVAN, M. ZIMMER & R. RICHARDS, *supra* note 21, at 78-80.

24. Baldus and Cole argue that "a general presumption should exist in favor of actual applicant data as a preselection basis of comparison. . . . [T]he unavailability of actual applicant data, or the possibility of distortion from the use of applicant flow data, can provide a basis for using a pool of potential applicants as a proxy for the people who would have applied under conditions of normal labor supply." P. 106 (footnote omitted). They offer guidelines for construction of such "proxy populations." Pp. 115-34.

ation of the . . . argument that those data may undercut a statistical analysis dependent upon hirings alone."[25] On this point, at least, the "man of statistics" will find that the Supreme Court's opinion passes muster.

But the "man of statistics" will be sorely troubled by how the Court used the data before it. It may be trite to say that a little knowledge is a dangerous thing, but in the case of the statistical reasoning in *Hazelwood* it is true enough. In suggesting that a high probability of sampling error tends to prove the absence of discrimination, the Court is doing what careful statisticians always warn against — trying to prove the "null hypothesis."[26] Under the classical theory of hypothesis testing, all that can be said is that the numbers do not compel us to reject the thought that the hiring process is free from racial discrimination. Of course, the failure to find something is usually a good indication that it is not there, but a statistical analysis of how a negative finding might "weaken" the government's case would require techniques that go beyond the classical methodology to which the Court refers.[27]

Moreover, even in its own terms, the Court's reasoning seems faulty. To see the problem, we must understand where the "rule" about two or three standard deviations comes from. The standard deviation is a measure of how widely varied a set of numbers is. (For those who like formulas, it is calculated by finding the mean of all the numbers, subtracting this mean from each number, squaring this difference for each number, adding all the squares together, dividing by size of the set, and finally extracting the square root of the resulting quantity.) When the disparity between the number of blacks actually hired and the number expected in a race-neutral process (without any sampling error) is measured in units of standard deviations, it is easy to deduce the probability that the observed disparity is the result of sampling error if certain conditions hold. For brevity, I shall not explain the details of this process. Suffice it to say that in situations where one has no idea in which direction the disparity will lie, a discrepancy of roughly two standard deviations in a large, randomly drawn sample implies that the probability that such a difference would arise by chance alone is no more than .05, or one

25. 433 U.S. at 313 n.21.

26. In a rare article finding "fundamental flaws" in *Hazelwood*, Smith and Abram make this point. Smith & Abram, *Quantitative Analysis and Proof of Employment Discrimination*, 1981 U. ILL. L. REV. 33, 52-53.

27. *See* notes 70-73 *infra* and accompanying text.

out of twenty. Three standard deviations corresponds to a probability of about .001, or one out of a thousand.

So the *Hazelwood* Court's reasoning comes down to this: With the City of St. Louis excluded from the labor market, the probability that so few blacks would be hired by reason of chance alone is larger than .05, and as long as this chance remains even slightly higher than this .05 level, the statistical evidence "weakens" the claim that hiring is improperly influenced by race. Yet, modern statisticians do not woodenly insist on a significance level of .05. Although the .05 level has become conventional in social science research, most thoughtful statisticians deplore the convention and urge that researchers state the probability level involved to permit the reader to reach his own conclusion about the significance of the result.[28] The fact that many social scientists feel that they should not claim to have discovered something new unless they can attach a probability value of less than .05 to the likelihood that they are merely observing sampling error hardly means that such small probability values are required in proving facts in civil cases.[29] A researcher may not wish to rush into print only to be contradicted by his colleagues when they attempt to replicate his results. He may wait to gather more data instead of putting his reputation on the line and perhaps causing others to spend time and money verifying tentative and misleading results. The concerns and values of social science, however, do not necessarily govern legal proceedings. In civil litigation, a less demanding more-probable-than-not standard is ordinarily employed. The Court is wrong in suggesting that a plaintiff does not make out a prima facie case under this standard unless the probability associated with sampling error is below .05.[30] Depending on the other evidence in the case, a much higher — or lower — value may suffice.[31]

Furthermore, even if the arbitrary choice of the .05 level were more defensible, the attorney familiar with *Statistical Proof of Dis-*

28. *E.g.*, D. MOORE, STATISTICS: CONCEPTS AND CONTROVERSIES 291-93 (1979); Skipper, Guenther & Nass, *The Sacredness of .05: A Note Concerning the Uses of Statistical Levels of Significance in Social Science*, in STATISTICAL ISSUES: A READER FOR THE BEHAVIORAL SCIENCES 141 (R. Kirk ed. 1972).

29. *See* Lempert, *Uncovering "Nondiscernible" Differences: Empirical Research and the Jury-Size Cases*, 73 MICH. L. REV. 644, 658-59 (1975).

30. *See* Smith & Abram, *supra* note 26, at 43-44.

31. Baldus and Cole also argue that the appropriate significance level will vary according to the type of discrimination case involved. P. 318. The only example they offer is a comparison of a challenge to a capital sentence and an attack on an employer's practices. If the significance level should differ in these two situations, it must be because the burden of persuasion — which turns on the relative costs of type I and type II errors, *see, e.g.*, Kaplan, *Decision Theory and the Fact-Finding Process*, 20 STAN. L. REV. 1065 (1968), — is a function of these costs. *Cf.* note 41 *infra* (discussing these types of mistakes).

crimination should wonder whether the Court was simply mistaken in concluding that the .05 test was not satisfied by the statistical evidence in *Hazelwood*. I stated earlier that two standard deviations correspond to a probability of .05 *if*, among other things, one has no reason to expect that the difference between the proportion for the labor pool and for the teachers hired will be in one direction as opposed to the other (a "two-tailed" test). In *Hazelwood*, however, the question is whether the employer is discriminating *against* a protected group. Therefore, the statistical problem in classical terms is to calculate how likely it is that the number of black teachers hired would be so much *less* than the number that race-neutral selection would produce in the absence of any sampling error.[32] Putting the question this way requires a "one-tailed" test and implies that it takes not two, but only 1.64 standard deviations to reach the .05 level. For the two-year period for which figures on hiring are described in the *Hazelwood* opinion, the number of standard deviations is less than two, as the Court states. But it is more than 1.64. By my calculation, it is 1.73, which corresponds to a probability of .04 of a sampling error. Consequently, one might well conclude that even at the demanding .05 level the Court arbitrarily selected, the chance that so few black teachers would be hired over the two-year period if selections really were independent of race is small enough to warrant rejecting the view that the school district did not discriminate against blacks.

I hope that this evaluation of the use of statistical methods in *Hazelwood* does not foster the impression that statistical analysis should be avoided at all costs. The Court performed poorly in *Hazelwood* not because it knew too much about statistical reasoning, but because it knew too little. Mathematical analysis cannot dictate answers to legal issues, but as Baldus and Cole urge, it can be a valuable aid in certain litigation. As the courts gain experience with statistical techniques in discrimination cases, they will learn to avoid

32. In technical jargon this is to say that a one-tailed test should be used in preference to the *Hazelwood* two-tailed test. A law clerk to Justice Stevens appears to have recognized the issue. *See* Hazelwood v. United States, 433 U.S. at 318 n.5 (dissenting opinion). Baldus and Cole avoid taking a stand on this issue. They write that "no strong conventions exist on the subject," but note that "statistics texts frequently recommend the use of a one-tailed test when the only question of interest is the likelihood of a difference in one direction . . . ." P. 307. They use two-tailed tests in their examples but suggest that "[s]ince there is no clear answer to this question, the most desirable approach is an awareness of the conceptual and practical differences between the two types of tests and a consistent use of the same type of test in similar cases whenever practical." P. 308. For a sampling of the social science literature on the propriety of one- versus two-tailed tests, see Jones, *Tests of Hypotheses: One-Sided vs. Two-Sided Alternatives*, in R. KIRK, *supra* note 28, at 276-90.

errors like those made in *Hazelwood*. In the meantime, the learning process is bound to be slow and, I fear, painful.

II. Where Do We Go from Here?

A. *Must We Think About Statistical Significance?*

In the hope of accelerating this learning process, I would like to canvass various alternatives to the hypothesis testing prospective adopted in *Hazelwood*.[33] Four of the five approaches that I shall enumerate are discussed in *Statistical Proof of Discrimination*, although the quality and depth of treatment varies widely.

One such alternative is simply not to bother inquiring into statistical significance at all. This is not an entirely frivolous suggestion. In *Hazelwood* itself, the Court noted that its calculations were "not intended to suggest that precise calculations of statistical significance are necessary in employing statistical proof . . . ."[34] There is also some academic support for this view.[35] Those of this persuasion who are also statistically sophisticated argue that "[w]hen the data comprise all the observations of the defendant's reward allocation process (i.e., the [whole] universe or population of observations), statistical inferences and tests of statistical significance are inappropriate."[36]

For anyone who appreciates Carlyle's quip that "I don't pretend to understand the Universe — it's a great deal bigger than I am,"[37] this "whole universe" argument fails. The numbers generated in a discrimination case describe only a sample of observations, but truly interesting conclusions concern a larger population. Statistical inference — the process of saying something intelligent about an entire population on the basis of sample data — is therefore unavoidable. Take the hiring process in *Hazelwood*. Certain teachers interviewed

---

33. The hypothesis testing in *Hazelwood* was presaged by a similar analysis (rejecting the null hypothesis) in Castaneda v. Partida, 430 U.S. 482 (1977). In *Castaneda*, the Court recognized that in speaking of how two or three standard deviations are necessary to establish statistical significance, it was simply stating a convention adopted in social science. 430 U.S. at 496 n.17. In *Hazelwood* the Court quoted from *Castaneda*, but it dropped *Castaneda*'s qualifying language.

34. 433 U.S. at 312 n.17.

35. Cohn, *supra* note 12, at 1304-07; Cohn, *On the Use of Statistics in Employment Discrimination Cases*, 55 Ind. L.J. 493, 494-99 (1980); *authorities cited at* p. 316 n.46. For contrary views see pp. 316-17; Shoben, *In Defense of Disparate Impact Analysis Under Title VII: A Reply to Dr. Cohn*, 55 Ind. L.J. 515 (1980); Smith & Abrams, *supra* note 26, at 42-43 (citing Freeman, *Availability, Goals and Achievements in Affirmative Action: An Economic Perspective*, in Perspectives on Availability 95, 110 (Equal Employment Advisory Council 1977)).

36. Cohn, *supra* note 12, at 1305.

37. D. Sciama, The Unity of the Universe 47 (1961) (quoting Carlyle).

for jobs, and the school district hired some of them. Those who went through this process should be thought of as a sample drawn from the population of potential applicants. As a first approximation, those whom the school district hired can be treated as a sample drawn from this sample of actual applicants. Of course, the successful interviewees were not hired at random, but if the characteristics legitimately considered in the hiring process are not correlated with race, then the probability model used by the Court is an appropriate one for estimating the probability that a sample selected on the basis of these criteria would contain so few black teachers. Surely, this probability (often called a p-value) is the kind of number that merits attention in determining whether a prima facie case of discrimination exists. A small enough p-value should lead us to think that the simple race-independent model of hiring describes the actual selection process poorly. We should then ask the defendant to provide a better model to rebut the charge of discrimination.[38] As such, the temptation to ignore the question of statistical significance by invoking the "whole universe" problem should be resisted.

B. *Unbridled Intuition*

Since statistical inferences must be made, one way or another, when quantitative evidence is introduced in discrimination cases, a method that is both logically defensible and intelligible to judges or juries should be employed. Elegant mathematical procedures are not the only possibility. Presented with a numerical disparity in group outcomes and the size of the sample giving rise to this statistic,

38. Cohn contends that "[r]elying on tests of statistical significance, based on the assumption that the unspecified determinants' effects on groups' outcomes are random, is a poor substitute for the proper modeling of the defendant's reward allocation process that would include all the relevant determinants of rewards." Cohn, *supra* note 12, at 1306. Speaking as a sociologist about academic studies, he is correct. To build the most accurate and plausible model, one should include all the important independent variables (taking into account problems that small sample size and multicollinearity may create). The plaintiff in a law suit, however, should not be required to build such an elaborate model if (1) the statistics derived from the (inevitably) small sample data indicate that the cruder race-independent model does not come close to explaining the apparent underrepresentation of blacks or some other protected class, *and* (2) it would be difficult and expensive to develop the full model, or there is ample qualitative evidence of discrimination. The second condition is important in light of the problem of "naked statistical evidence." *See* Kaye, Book Review, 89 YALE L.J. 601, 603 (1980).

Cohn's argument against statistical inference is vital to his claim that *Statistical Proof of Discrimination* "fails to draw a sharp distinction between [causal inferences and statistical inferences] and misleads the reader by suggesting that statistical significance tests indicate the strength of the causal inference that the defendant's discrimination caused group differences in outcome." Cohn, *supra* note 12, at 1302. This criticism seems overdrawn in view of Baldus's and Cole's admonition that "the test of significance speaks only to whether there is some difference in the universe . . . . [S]tatistical tests can tell us nothing, directly, about the cause of an adverse impact . . . . [T]he test provides no basis for assigning a precise probability to possible causes." P. 320. *But see* note 42 *infra*.

a fact-finder could simply try to intuit how likely it is that this difference results from something other than the luck of the draw. But why should we not use valid statistical tools to provide the fact-finder with additional information that is pertinent to his assessment of this likelihood?[39]

C. *P-Values*

With respect to the formal statistical methods that might be employed, I have already argued that testing the null hypothesis at a fixed significance level such as .05, as in *Hazelwood*, is inadvisable.[40] I would only add that if strict hypothesis testing is used, some indication of the so-called "power" of the test or related quantities should be given.[41] Baldus and Cole favor the more flexible procedure of

39. To be sure, unless we were to modify radically our legal procedures, most of the evidence in a discrimination case will have to be weighed intuitively, and in the end the statistical evidence may have to be judged along with the qualitative evidence in some intuitive fashion. Nonetheless, statistical techniques can aid in this intuitive process. *Cf.* p. 317 ("one is in a better position to assess the long-run effects of a rule with the help of statistical tests than without").

40. *See* note 31 *supra* and accompanying text.

41. The power of a test is the probability that it correctly rejects the null hypothesis. It is the complement of the probability that the test will accept the null hypothesis when that hypothesis is false. This latter probability is — or should be — of great interest in formulating or evaluating an hypothesis test. If it is high, the test rarely will detect the defendant's discrimination. If it is low, the test is sensitive to the presence of discrimination — it is powerful.

Unfortunately, the relationship between this probability and a test's significance level is complex, and it tends to be ignored in discussions of statistical proof of discrimination. *E.g.*, C. SULLIVAN, M. ZIMMER & R. RICHARDS, *supra* note 21; Braun, *supra* note 21. *But see* Dawson, *Are Statisticians Being Fair to Employment Discrimination Plaintiffs*, 21 JURIMETRICS J. 1 (1980) (examining the power of hypothesis tests that use chi-square as the test statistic). Baldus and Cole do not address the topic (presumably because they argue against formal hypothesis tests in the first place), and except in a footnote at page 291, they make no explicit mention of this probability.

Accordingly, a brief description of the issue may be in order. In designing or adopting a hypothesis test, one should keep two factors in mind: the probability that the test will reject the null hypothesis when that hypothesis is in fact true, and the probability that the test will fail to reject the null hypothesis even though that hypothesis is false. The first probability measures the chance of what is variously called a type I error, a false rejection, a false positive or a false alarm. Customarily denoted by the Greek letter α, it is nothing other than the significance level of the test. The second probability quantifies the chance of the opposite type of error — a type II error, a false acceptance, a false negative, or an undetected signal. Customarily denoted by β, its value is typically far harder to discern. It is *not*, as one might think, $1-\alpha$.

The hypothesis test outlined in *Hazelwood* illustrates the situation. In always rejecting the null hypothesis of no discrimination at a significance level of $\alpha=.05$, the Court will erroneously reject this hypothesis five percent of the time, since five percent of the cases triggering rejection will occur due to chance alone, and the null hypothesis asserts the existence of such chance results. On the other hand, suppose the null hypothesis is false. What is the probability that the test will not detect this fact? It is not $1-\alpha=.95$. The alternative to the null hypothesis in the Court's test is that the chance that each black applicant will be hired differs from .057, the proportion of blacks in the relevant labor market. It could be .056, .999, .030303, or any other number other than .057 between zero and one. For each such number — and the possibilities are infinite — there is some probability that the test will still accept the null hypothesis. Imag-

presenting the p-value, that is, the probability that the measured discrepancy is merely a chance fluctuation, one that would disappear in an examination of the defendant's long-run behavior.[42] The *Hazelwood* Court, it will be recalled, compared the "observed number" of black teachers hired over a two-year period (15) to the "expected number" (23) derived from a probabilistic model in which the chance of hiring a black teacher was .057 in each instance. The Court took the .057 figure from the areawide data on the proportion of black teachers in the suburban area, and it used the sample size (the number of teachers, 405, who actually applied for positions in the two-year period) to conclude (using the normal approximation to the binomial distribution) that the p-value fell below the .05 significance level (using a two-tailed test). Had the Court followed Baldus's and Cole's recommendation, and merely stated the p-value, it would have noted that under the posited model of no discrimination the probability that no more than fifteen blacks would have been hired is only .042.[43] The calculation of such a quantity, Baldus and Cole suggest, is ordinarily the task of the expert, but the evaluation of this number — which is an aspect of the assessment of the probative force of the statistical evidence — lies within the peculiar province of the judge or jury.[44]

---

ine, for example, that each black applicant stood a .047 chance of being hired. The resulting probability of a gap as large or larger than the difference between the associated expected number (19) and the observed number (15) would then be about .14. Thus, there is a .86 probability that the Court's hypothesis test would not detect any discrimination in Hazelwood's hiring practices even if those procedures reduced each black teacher's chance of being hired from .057 to .047 (a reduction of 18 percent, depriving ten out of every thousand black applicants of positions). When it comes to detecting this degree of discrimination, the Court's test is not very powerful. The chance of a false alarm is small (.05), but the chance of an undetected signal is large (.86).

For elementary, but more complete descriptions of the power of hypothesis tests, *see, e.g.*, Y. CHOU, STATISTICAL ANALYSIS 283-96 (2d ed. 1975); B. LINDGREN, G. MCELRATH & D. BERRY, INTRODUCTION TO PROBABILITY AND STATISTICS 186-93 (4th ed. 1978). For a revealing but rather mathematical treatment of hypothesis testing from the standpoint of signal detection theory, see J. MELSA & D. COHN, DECISION AND ESTIMATION THEORY 21-53 (1978).

42. This characterization is oversimplified. The p-value is computed on the assumption that the null hypothesis is true. The null hypothesis entails a particular probability model with specific parameters. *See* note 59 *infra*. In *Hazelwood*, for example, the Court assumed that selection was a Bernoulli process with the binomial parameter $\theta = .057$. Consequently, a very small p-value suggests that either the parameter has some other value ($\theta \neq .057$) or that the form of the model is wrong (selection of teachers is not a Bernoulli process). *Cf.* pp. 290-93 (discussing the use of p-values in disparate treatment cases to reject a Bernoulli model in favor of some other model involving "legitimate selection criteria" or "defendant's bias," and in this way linking "causal inference" with "statistical inference").

43. This is the p-value that would be used implicitly in a one-tailed test. *See* note 32 *supra*. For a two-tailed test, one would compare, in effect, the probability that the number of blacks hired would be less than 16 or greater than 30 with the preset value α. This two-sided p-value is .084, which is greater than $\alpha = .05$ — the result mistakenly emphasized by the Court.

44. P. 308. *Cf.* R.A. FISHER (1955), *quoted in* A.W.F. EDWARDS, LIKELIHOOD: AN ACCOUNT OF THE STATISTICAL CONCEPT OF *LIKELIHOOD* AND ITS APPLICATION TO SCIENTIFIC

On pragmatic grounds, a few commentators question the desirability of producing p-values. Kairys, Kadane, and Lehoczky have written that it "involves complicated calculations resulting in answers that are difficult to visualize and evaluate," and they complain that "[e]ven with moderate sample sizes, small disparities result in very low probabilities."[45] Yet, the calculation of the p-values is not difficult, and the concept of sampling error or measurement noise is not beyond a court's comprehension. That with moderate or large samples even trivial differences can have small p-values and therefore appear statistically significant should not trick most fact-finders into thinking that such a disparity is also legally significant.[46] This concern does underscore the admonition that p-values should not be considered in a vacuum, but the courts are not likely to shut their eyes to the possibility that although the group difference is statistically significant, the degree of discrimination is itself de minimis. For these reasons, the use of p-values as Baldus and Cole recommend seems superior to the alternatives of doing nothing in the way of formal inference or of deciding according to rigid, classical hypothesis tests.

D. *Prediction Intervals*

Still, there is one more classical technique that promises to convey yet more information to assist the court in arriving at an ultimately intuitive assessment of the evidence. It is most easily described in the context of an example, so I shall return to the *Hazelwood* case once more. Both the computation of the p-value and the standard deviation "rule," or hypothesis test, use a "point estimate" of twenty-three for the "expected number" of blacks hired. In other words, these methods suppose that if repeated samples of 405 applicants were to be assessed under the Court's simple probability model and if these samples were to be pooled, then the proportion of black teachers in the ensuing collection of successful applicants would ap-

---

INFERENCE, at v (1972) ("We [statisticians] have the duty of formulating, of summarizing, 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"). Arguably, the p-value could be better assessed if a statement of the power of an hypothesis test that treats the observed disparity as (barely) significant were provided. *See* note 41 *supra*.

45. Kairys, Kadane & Lehoczky, *Jury Representativeness: Mandate for Multiple Source Lists*, 65 CALIF. L. REV. 776, 794 (1977). These objections are offered in the context of jury selection cases. Kairys, Kadane, and Lehoczky note that "[t]he problem of sample size is not so acute in employment discrimination cases" because "the sample size is the number of people hired and is typically small." *Id.* at 794 n.101.

46. Baldus and Cole caution against this erroneous interpretation of statistical significance. *See* pp. 317-20.

proach 23/405. In the limit (as the pooled sample grew to engulf the entire population of teachers in the relevant labor market[47]), the observed proportion would *be* the population proportion of 23/405. But probability theory permits us to provide more than a point estimate of the number of successful blacks. Given the Court's probability model, it is easy enough to compute an "interval estimate." Figure 1 presents several such prediction intervals.[48]

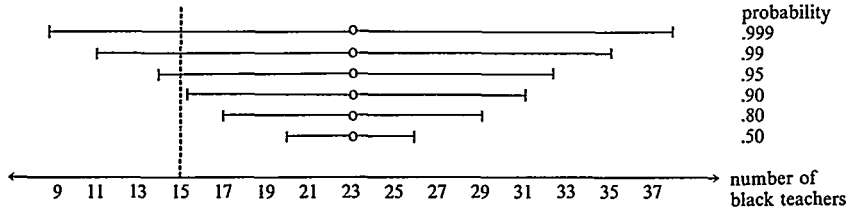


Figure 1. Prediction intervals for the number of black teachers hired out of 405 applicants assuming that the probability of hiring a black teacher is .057 in each instance. The observed number of black teachers hired is 15, which lies outside the .90 prediction interval but inside the .95 prediction interval. The disparity is therefore significant at the .10 level, but not at the .05 level, using a two-tail test.

All the intervals are centered about the previous point estimate of twenty-three, and the wider the interval, the more likely it is that the number of blacks actually hired from a group of 405 applicants will be included in the predicted interval. For example, the probability that the number of blacks will be between eighteen and twenty-eight is .75. In contrast, if we choose a wider prediction interval, say one that runs between sixteen and thirty, we can be more confident that the number of blacks hired will fall within the interval. Specifically, the probability that this range of outcomes will contain the observed outcome is .90. Saying the same thing another way, we predict that the observed number will lie outside the interval [16,30] in only one sample of 405 out of every ten.[49] The fact that the observed number

47. Since the pool of potential applicants changes as some teachers enter the market while others leave, the population size is infinite.

48. Prediction intervals are akin to, but conceptually distinct from "confidence intervals." *Statistical Proof of Discrimination* gives a clear explanation of confidence intervals and their relation to p-values and significance tests. Pp. 310-13.

49. The perceptive reader will recognize from this example that the confidence level p is intimately related to the significance level α. Namely, $p = 1-\alpha$. Consequently, interval estimates can readily be used to perform significance tests.

does fall just outside this range (it is 15) reveals that either (1) we just happen to have before us one of the unusual cases that arise ten percent of the time, or (2) the Court's no-discrimination model is a poor description of the long-run characteristics of the hiring process.

This type of presentation is more elaborate than a terse statement that the probability of picking no more than fifteen blacks on 405 tries would be .042 if the probability on each try were .057. While its import is essentially the same as that of the p-value,[50] it should be useful in helping a court visualize the statistical issue and therefore should be made available to supplement the p-value itself.[51]

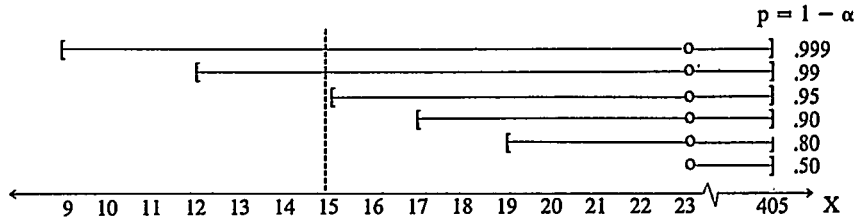
III. NONCLASSICAL ANALYSES

A. *Likelihood*

Although the full conceptual apparatus of p-values, prediction intervals, and significance tests was not fashioned until the late 1920s,[52] these techniques have already earned the sobriquet "classical." Contemporary statisticians all agree that these techniques are of some use in evaluating limited observations of long-run phenomena. At the same time, many prominent statisticians believe that other methods deal more effectively with the problem of statistical inference. To appreciate these competing approaches, we must be clear about the nature of the problem. As one leading text on mathematical statistics explains:

A problem of statistical inference or, more simply, a statistics problem

50. Since the .042 p-value is one-sided (it measures the probability that the number of blacks selected would not exceed fifteen if each black's chance of selection were .057), it is more closely connected with one-sided prediction intervals. I have displayed two-sided prediction intervals in the text only because it seems to me that these are slightly easier to grasp. The corresponding one-sided prediction intervals are shown below:



Since 15 is outside the prediction interval for p = .95, the small number of blacks hired is significant at the .05 level (using the one-tailed test).

51. *Cf.* p. 310 ("the confidence interval answers a broader question than that addressed by tests of significance"); Natrella, *The Relation Between Confidence Intervals and Tests of Significance*, 14 AM. STATISTICIAN 20 (1960) (advocating more widespread use of confidence intervals in statistical studies).

52. *See, e.g.*, Dudycha, *Behavioral Statistics: An Historical Perspective*, in R. KIRK, *supra* note 28, at 2, 21-24.

> is a problem in which data that have been generated in accordance with some unknown probability distribution must be analyzed and some type of inference about the unknown distribution must be made. In other words, in a statistics problem there are two or more probability distributions which might have generated some given experimental data. In some problems there could be an infinite number of different possible distributions which might have generated the data. By analyzing the data, we attempt to learn about the unknown distribution, to make some inference about certain properties of the distribution, and to determine the relative likelihood that each possible distribution is actually the correct one.[53]

This is precisely what the Court was attempting to do in *Hazelwood*. Looking at sample data (involving 405 applicants over a two-year period), the Court asked whether it was reasonably likely that the unknown distribution giving rise to the data (summarized by the sample statistic that fifteen blacks were hired) was a so-called binomial distribution whose parameter was .057.[54] Because the probability of the observed number's being generated by this distribution was a bit more than .05 (calculated by a two-tailed test), the Court suggested that its tentative guess about the nature and details of the unknown distribution was pretty good.

Upon reflection, however, this reasoning seems to leave out a crucial ingredient. It certainly tells us something about how well the binomial no-discrimination model fits the data, but it reveals nothing about the accuracy of other models. If, under a revised model of the hiring process, the probability of the observed statistic would be higher than that calculated under the Court's version, then this other model would seem to emerge as a more likely prospect for the unknown distribution. While one must be wary of "overfitting" a model to the data, a systematic way to look at the relative likelihood of various hypotheses about the unknown distribution is available.

This procedure requires the construction of a "relative likelihood function," a task that is not as complicated as it might sound. Let us denote the unknown parameter of the posited binomial distribution by the Greek letter θ, and let X stand for the number of black teachers actually hired.[55] So far, we have treated X as a *variable* whose

---

53. M. DEGROOT, PROBABILITY AND STATISTICS 257 (1975).

54. A binomial distribution describes the probability of the number of "successes" for some number of trials, where the outcome of any one trial is independent of the outcome of any other trial. The probability of a success on any particular trial is a "parameter" of this binomial distribution. *See, e.g., id.* at 201. The probabilities of the various possible numbers of heads obtained by flipping a fair coin ten times are given by a binomial distribution with the parameters ten (for the number or trials, or the sample size) and .5 (for the probability of a success on each trial).

55. The Court's approach tests the "null hypothesis' that $\theta = .057$ against the "alternative

value is determined by the luck of the draw — that is, by random sampling from an infinite population characterized by the binomial distribution whose parameter θ is .057. Now we treat X as *fixed* — the school district hired fifteen blacks — but we think of θ as a variable that can take on values between zero and one. For each such value of θ, there is some probability $Pr(15;\theta)$ of obtaining the statistic X=15 from the sample of 405 applicants. One special value of θ, which we designate $\hat{\theta}$, maximizes this probability $Pr(15;\theta)$. That is, the probability of observing X=15 is highest when $\theta = \hat{\theta}$. It turns out, although I shall not prove it, that this maximum likelihood estimator of θ is $\hat{\theta} = 15/405$.

All this talk of θ's, $\hat{\theta}$'s, and related probabilities may sound complicated, but the idea expressed should be intuitively plausible. If we flip a possibly unfair coin ten times, knowing nothing about how the coin is weighted, and we observe a total of three heads, it is more likely that the coin has a probability $\theta = 3/10$ of coming up heads on each toss than that the coin is weighted in some other way.[56] In this case, $\hat{\theta}$ would be 3/10. Since $\hat{\theta}$ maximizes $Pr(15;\theta)$, it can serve as a standard against which to judge other hypotheses about the true value of θ. Suppose, for instance, someone exclaimed that in the *Hazelwood* case θ was not $\hat{\theta} = 15/405$, but some larger number, say 23/405. We might respond by designating this newly suggested value of θ as θ_1 and calculating $Pr(15;\theta_1)/Pr(15;\hat{\theta})$. The number would tell us how likely it is that $\theta = \theta_1 = 23/405$ relative to the more likely possibility that $\theta = \hat{\theta} = 15/405$. Calling this ratio the relative likelihood of θ_1 and denoting it by $R(15;\theta_1)$, it is not hard to show that $R(15;\theta_1) = .19$. In light of the observed value of X, the hypothesis that $\theta = \theta_1 = 23/405$ is about a fifth as likely as the hypothesis that $\theta = \hat{\theta} = 15/405$. We can now repeat this procedure for θ_2 = some other possible value of θ — for θ_3, and so on. Doggedly, the statistician (or his computer) finds the values of the relative likelihood function $R(X;\theta)$. Figure 2 shows this function for the *Hazelwood* situation. Scanning such a graph should give the factfinder a feeling for the plausible range of the true value of θ.

---

hypothesis" that $\theta \neq .057$. But, as the test is implemented by the Court, we learn nothing about the probability of obtaining the sample statistic under this alternative hypothesis. Computing this probability is no easy task. An entire family of binomial distributions in which the parameter θ takes on all the values (except .057) between zero and one must be considered. *See* note 41 *supra*.

56. Anyone inclined to say that the coin is not likely to be this biased and probably is characterized by a θ closer to .5, is almost surely being influenced by his prior beliefs about the prevalence of weighted coins. Likelihood methods, like classical inferential techniques, ignore prior beliefs. The accepted statistical method for incorporating prior beliefs into the inference problem is described at note 67 *infra* and accompanying text.

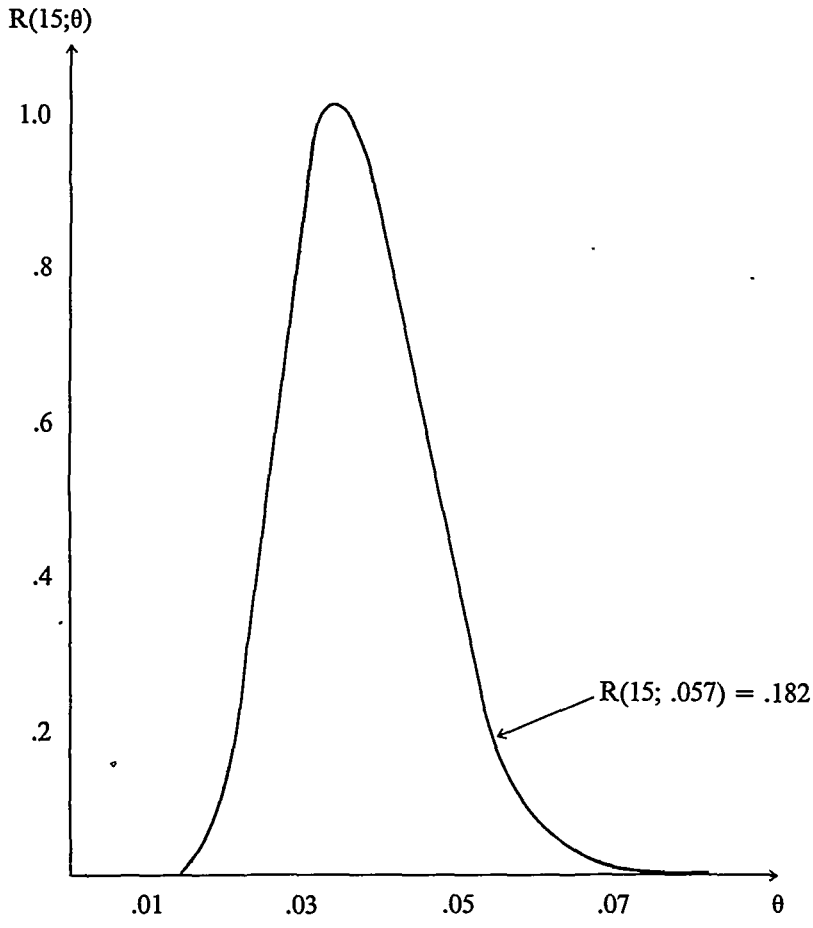


Figure 2. Relative likelihood function R(X;θ) for selecting X = 15 blacks out of 405 applicants where the probability of selecting each black applicant is θ. The likelihood that θ = .057 (which corresponds to the *Hazelwood* Court's no-discrimination hypothesis) is about 18% of the likelihood that θ = .037 (which is the most likely value of θ looking only to the number of blacks hired).

It bears emphasizing that the statistical analysis merely produces the picture. Defining the plausible range is not a mathematical operation.[57] That is a question for a judge or jury. The purpose of generating a graph of the relative likelihood function is merely to give

57. Likelihood intervals can be defined to capture those values of θ that yield relative likelihoods above a fixed amount. *See, e.g.*, J. KALBFLEISCH, PROBABILITY AND STATISTICAL

the factfinder more guidance than the classical p-value affords him. Anyone who looks carefully at Figure 2 must wonder about the *Hazelwood* Court's suggestion that .057 is a convincing value for θ. If the hiring process involves independent trials as the Court's null hypothesis presupposes,[58] then it is hard to believe (on a preponderance of the evidence standard, looking solely to the statistical evidence) that the probability that a black applicant will be hired is as high as .057. And, if it is lower than .057, black applicants do not have the same chance of being hired as white applicants do.[59]

I fear that much of this likelihood analysis will seem confusing — too abstract, or at the other extreme, too mired in the details of a messy example.[60] I have devoted substantial space to illustrating the rudiments of the likelihood theory because I think it has some potential for use in discrimination litigation. Admittedly, it is not nearly so familiar a way of thinking about the problem of statistical inference as are the classical methods that Baldus and Cole so ably discuss. Yet, any statistician who has seriously studied the fundamentals of inference should admit that the likelihood function contains all the information that the statistical findings can convey. It forces the fact-finder to confront all the hypotheses concerning the parameters of a probability distribution, and it avoids the arbitrary character of hypothesis testing at a uniform significance level. I would not say that it should replace p-values and prediction intervals (and their close cousins, confidence intervals), but it can supplement these standard methods.

B. *Bayesian Inference*

At this point, it might be wise to look back over the territory that we have traversed. We have seen (or, rather, I have asserted) that within the framework of a statistical model, the likelihood function contains all the information that the data provide concerning the rel-

INFERENCE II 22-27 (1979). But, as with prediction or confidence intervals, what threshold figure to use is not a question that mathematics can answer. *See* note 44 *supra*.

58. Note 38 defends the general use of the binomial distribution model where the question is whether the disparity in group outcomes is not only practically important, but also statistically reliable enough to make out a prima facie case of discrimination.

59. Perhaps they should not have the same chance. Maybe white applicants are generally more qualified than their black counterparts. But that is a point that the school district could raise — with the aid of a more sophisticated statistical model of the hiring process if need be — to rebut the prima facie case of discrimination. *See* note 38 *supra*.

60. For a relatively simple description of likelihood theory using other examples, see Sprott & Kalbfleisch, *Use of the Likelihood Function in Inference*, 64 PSYCH. BULL. 15 (1965). A.W.F. Edwards, *supra* note 44, gives a philosophically oriented survey of likelihood theory. For more detailed mathematical presentations, see, *e.g.*, J. KALBFLEISCH, *supra* note 57; D. FRASER, THE STRUCTURE OF INFERENCE 185-88, 295-316 (1968).

ative merits of the possible hypotheses. The prediction interval states a range within which the data would be expected to fall if the nondiscrimination version of the model applies. The p-value gives the probability that the data would be as observed in the specific version of this same model of the selection or allocation process that entails no discrimination. Oddly, none of these statistical constructs tells us what we really want to extract from the statistical evidence in a lawsuit: the probability, computed in light of the sample data, that the defendant's selection or allocation process involves disparate treatment or impact. Thus, in the *Hazelwood* situation, we have examined such things as $Pr(X \leq 15|\theta = .057)$ — the probability that X, the number of black teachers hired out of 405 applicants, would be fifteen or fewer, given that each black applicant had the same .057 chance of being hired — and $R(X;\theta)$ — the relative likelihood that θ has various values, including but not limited to .057.[61] As yet, we have exhibited no calculations of $Pr(\theta = .057|X = 15)$ — the probability, conditioned on the observation that the school district hired fifteen blacks, that each black applicant had the same .057 chance of being hired.

Neither likelihood methods nor classical techniques of inference can ever produce this figure. Yet, there is a rich body of statistical theory that permits such calculations. It goes by the name Bayesian inference.[62] As is true of likelihood methods, it is not used as widely as the classical theories, but it is becoming increasingly influential.[63]

Bayesian inference employs likelihood ideas in a distinctive way. It uses the likelihood function[64] to convert a "prior" probability dis-

61. This example may clarify the point that the important quantities in classical as well as likelihood theories are computed within the framework of a statistical model. As I have previously noted, *see* note 42 *supra*, that model in this instance is called a Bernoulli model in which the probability of a "success" on each trial is some fixed number θ, which is called the parameter of the model. *See* note 54 *supra*. Classical calculations (with the important exception of the power function mentioned in note 41 *supra*) take the value of this parameter to be fixed (for example, at .057). This version of the general model with the parameter so specified is what is meant by the "null hypothesis." Likelihood methods presuppose the same general model but treat the parameter θ as a variable and thereby permit the decision-maker to assess the innumerable versions of the model ignored by classical calculations of p-values and prediction intervals.

62. It should not be confused with the Bayesian, or subjective interpretation of probability (on which it builds). *See* Kaye, *The Laws of Probability and the Law of the Land*, 47 U. CHI. L. REV. 34, 50-52 n.56 (1979). This article lists several texts that explain Bayesian inference. *See generally* M. DEGROOT, *supra* note 53; J. KALBFLEISCH, *supra* note 57, at 288-94; B. LINDGREN, G. MCELRATH & D. BERRY, *supra* note 41, at 219-32.

63. *See, e.g.*, Schum, *A Review of a Case Against Blaise Pascal and His Heirs*, 77 MICH. L. REV. 446, 468 (1979).

64. The likelihood function $L(X;\theta)$ differs only slightly from the relative likelihood function $R(X;\theta)$. The latter is but a special case of the former.

tribution that characterizes an observer's belief[65] (about a population parameter such as θ) into a "posterior" distribution that takes account of the sample observations. For example, in the case of the coin that came up heads three times out of ten,[66] I asserted that the most likely value of a head appearing on each independent toss was 3/10. This number is the maximum likelihood estimate of θ, the extent to which the coin is weighted, in light of the limited data available. But if there were some prior reason to believe that the coin was in reality evenly balanced (perhaps the owner of the coin, a trustworthy soul, assured us that it is not a trick coin), we might be troubled by the idea of embracing 3/10 as our best estimate of θ. The 3/10 figure is plainly relevant, but must it be determinative? Bayesian analysis uses the sample data — the outcomes of the ten tosses — to revise the prior belief. Depending on the strength of the initial view that the coin is evenly balanced, a Bayesian would arrive at a point estimate that would put θ somewhere between .3 and .5.

The formal scheme for prescribing the impact of sample data on a prior distribution *could* be adapted to legal proceedings, although not as easily in discrimination cases[67] as in certain other contexts.[68] Baldus and Cole toy with the idea but curtly dismiss it, partly because "courts are even less familiar with Bayesian methods than they are with the methods of classical statistics."[69] Although it is difficult

65. Those who object in principle to Bayesian methods usually deny that it makes sense to take a probability distribution as characterizing belief. *See, e.g.*, G. SHAFER, A THEORY OF EVIDENCE (1976); Brilmayer & Kornhauser, *Review: Quantitative Methods and Legal Decisions*, 46 U. CHI. L. REV. 116 (1978). Of course, Bayes's rule can work with "objective" prior probability distributions instead of subjective ones. *See* J. KALBFLEISCH, *supra* note 57, at 291; Kaye, *supra* note 62.

66. *See* note 54 *supra* and accompanying text.

67. Sketching how one might proceed should convey some of the flavor of the difficulty of institutionalizing Bayesian inference in the law. In a case like *Hazelwood*, where the plaintiff alleged racial discrimination in hiring, the judge or jury could consider initially the plaintiff's nonstatistical evidence — testimony about the opportunities for discriminatory decisions, remarks of personnel managers evincing racial bias, and the like — as well as figures on the proportion of blacks in the relevant labor market or applicant pool. Under one approach, the fact-finder could then draw a curve on a chart, a curve that would reach its peak at the fact-finder's best estimate of θ and that would spread out in accordance with the fact-finder's confidence in this estimate. A statistician could derive the likelihood function from the sample data and apply it to this prior distribution via Bayes's rule to generate the posterior distribution. This distribution would supply the probability that θ is less than the proportion of blacks in the relevant population. A less confining elaboration of this procedure is given in the text accompanying note 70 *infra*.

68. *See* Finkelstein & Fairley, *A Bayesian Approach to Identification Evidence*, 83 HARV. L. REV. 489 (1970); Fairley, *Probability Analysis of Identification Evidence*, 2 J. LEGAL STUD. 1205 (1973); Lindley, *A Problem in Forensic Science*, 64 BIOMETRIKA 207 (1977).

69. P. 304 n.32. After a brief allusion to the "lively debate [in the Harvard Law Review] on the applicability of Bayesian methods to legal questions" they add that "[w]e know of no court that has applied Bayesian methods to an evidentiary problem." *Id. But see* Arizona v. Wagner, No. DR122023 (Super. Ct., Maricopa County, Ariz. 1980); Everett v. Everett, No. D-

to treat this as a serious argument,[70] Baldus and Cole also comment that "there is no judicially acceptable way we know of to quantify . . . the prior distribution" (p. 304). Presumably, they have in mind Professor Tribe's forceful critique of "trial by mathematics."[71] However, Tribe's arguments are not all of the same high caliber, and the most penetrating are blunted or sidestepped by circumspect procedures for exposing judges or juries to Bayesian logic.[72] Such procedures would not dictate the choice of the final distribution or the result in the case; they would merely demonstrate the probative force of the sample data by displaying the effect of the data across a panoply of prior distributions. A fact-finder would be free to start with any distribution that reflected his prior estimate of the unknown parameter, or to remain uncommitted to any specific estimate. He could merely see how strongly the likelihood function for the sample data affects various prior probability statements. For example, starting from the agnostic premise in *Hazelwood* that θ is no more likely to have one value between 0 and 1 than any other leads to the conclusion that, in light of the observed hiring rate, the probability that the true value of θ is less than .057 (*i.e.*, that selection is discriminatory in terms of the Court's model) is a bit over 95 percent.

I make these points primarily to give Bayesian methodology its fair hearing. I, too, would not urge its implementation in discrimination litigation. Even if the flexible use of Bayesian calculations would be no more confusing to the fact-finder than the alternatives, it would not provide that much additional guidance to the fact-finder. Some resort to intuition still would be essential, and the marginal guidance from seeing the results on a cross section of distributions does not seem large enough to justify implementing the procedure.

CONCLUSION

Lawyers and statisticians make strange bedfellows. Both are skilled (or should be) in the analysis of evidence. Yet, the two pro-

850-370 (Super. Ct., Los Angeles County, Cal. 1981); Ellman & Kaye, *Probabilities and Proof: Can HLA and Blood Tests Prove Paternity*, 54 N.Y.U. L. REV. 1131 (1979); Kaye, *supra* note 62, at 34 n.5 (referring to other cases applying Bayesian methods to an evidentiary problem, although not necessarily in a considered or deliberate way).

70. The argument from unfamiliarity is weak at best. *See* Ellman & Kaye, *supra* note 69, at 1158. It is especially troublesome in a work that aspires to be a response to "[t]he challenge . . . to weigh intelligently the strengths and weaknesses of the available methods and to use them creatively to focus on what is at issue under the substantive law." P.3.

71. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329 (1971).

72. Ellman & Kaye, *supra* note 69, at 1154-57.

fessions often reach apparently divergent conclusions about what counts as "evidence" and about what inferences can properly be drawn from such evidence.[73] Indeed, the legal profession has long resisted the allure of quantified methods of proof, partly out of the realization that values other than accuracy in factfinding are sometimes central to the proper resolution of legal disputes.[74] Nevertheless, the ability to ascertain the true state of affairs is important, and there are instances in which statistical data can enhance the quality of judicial decision-making. Especially in discrimination litigation, quantitative analysis is becoming, if not *de rigeur*, at least an accepted method of proof. *Statistical Proof of Discrimination* represents a sustained effort to make this kind of analysis understandable and useful to attorneys and courts involved in such litigation. There is surely more to be said on the subject,[75] but *Statistical Proof* is an excellent contribution to a burgeoning cross-disciplinary literature. Read in conjunction with a good elementary statistics textbook,[76] it can be an invaluable guide to the perplexed.

---

73. L.J. COHEN, THE PROBABLE AND THE PROVABLE (1977).

74. *See* Tribe, *supra* note 71.

75. The National Academy of Sciences' Committee on National Statistics is studying the use of statistics in litigation, and the National Science Foundation has funded a University of Minnesota study along these same lines. In addition, an exchange of papers on many of the issues canvassed in this Review as well as some related ones is scheduled for publication in the Journal of the American Statistical Association.

76. Nicely written works that use a minimum of mathematics include G. KIMBLE, HOW TO USE (AND MISUSE) STATISTICS (1978); D. MOORE, *supra* note 28; J. PHILLIPS, JR., STATISTICAL THINKING: A STRUCTURAL APPROACH (1973); E. WILLEMSON, UNDERSTANDING STATISTICAL REASONING: HOW TO EVALUATE RESEARCH LITERATURE IN THE BEHAVORIAL SCIENCES (1974).