Statistical Proof of Employment Discrimination

Sandy L. Zabell *Northwestern University*

Title VII of the Civil Rights Act of 1964 prohibits discrimination in employment on the basis of race, color, religion, sex, or national origin. In part because of the passage of the Civil Rights Act, and in part because of changes in the social climate, discrimination of the type prohibited under the Act is usually not as blatant as it once was, and proof of its existence is correspondingly much harder than before the Act's passage. During the last decade the courts have turned increasingly to the methods of statistics to assist them in assessing the existence and extent of discrimination.

The U.S. Supreme Court has played a key role in the judicial adoption of statistical methodologies to prove employment discrimination. Although statistical proof began to appear sporadically in the lower courts almost immediately after passage of the Civil Rights Act, during the past two decades the discussions of statistical proof in certain key Supreme Court decisions has played an important part in the willingness of the lower courts to accept such proof and in their increasing reliance on it.

DISPARATE TREATMENT AND DISPARATE IMPACT

Employment discrimination cases fall into two broad categories: *disparate treatment* and *disparate impact.* In a disparate treatment case, the issue is whether an employer has intentionally discriminated against someone. In such cases statistics can be introduced as circumstantial evidence to establish a "pattern and practice" of discrimination (as was done in *Hazelwood,* described below). In *Griggs* v. *Duke Power Company* (401 U.S. 424, 1971), however, the Supreme Court ruled that employment practices called "facially neutral," such as tests for hiring and promotion, were also discriminatory under Title VII if they had an adverse or disparate impact on a protected class. Even when an employer does not intend to discriminate, if the employment practices result in the hiring or promotion of protected classes of individuals at a substantially lower rate, a prima facie case has been made that the employer is in violation of the Act. By divorcing *outcome* from *intent,* the Supreme Court created an essentially statistical category of discrimination, and that in turn made the use of statistical proof inevitable.

Proving that a facially neutral employment practice has a disparate impact does not necessarily suffice to prove prohibited discrimination. Under the "business necessity" defense, an employer can defend a practice by showing that it has a legitimate business purpose. If, for example, it can be shown that an employment test is testing for a legitimate job skill, then this can serve as a defense. Validating an employment test (showing that it does test for such a skill) is also an area where statistics can be used, but, as a practical matter, it may often be hard for an employer to do such validation in an adversarial setting, and thus the demonstration of a substantial statistical disparity will often suffice to decide a case.

HAZELWOOD

Although the Supreme Court had on several earlier occasions pointed to statistics as a valuable tool in employment discrimination cases, its use of statistical tests of significance in *Hazelwood School District* v. *United States* (433 U.S. 299, 1977) gave the formal methods of statistical proof a stamp of approval they had not enjoyed previously. Since *Hazelwood* the lower courts have felt free to use statistical methodology in deciding employment discrimination cases.

In *Hazelwood,* the U.S. government brought suit against the Hazelwood School District, alleging that it had engaged in a "pattern or practice" of employment discrimination in violation of Title VII. Hazelwood is a largely rural school district in the northern part of St. Louis County, with few blacks in its student body and few (under 2%) black teachers on its staff. The government based its statistical case on the fact that the percentage of black teachers in the district was substantially smaller than in the surrounding area of St. Louis County and the city of St. Louis.

The district court (where the case was first tried) compared the percentage of black *teachers* in the district with that of black *students,* noted that they were roughly the same, and ruled in favor of the Hazelwood School District. When the case was reviewed, the Federal Court of Appeals ruled that the comparison of teachers with students was irrelevant and that the appropriate comparison was the percentage of black teachers in the *school district* with that of the black teachers in the relevant *labor market* (just as the government had argued), and reversed the lower court decision. The case was then reviewed by the Supreme Court.

The Supreme Court's decision notes that of 405 teachers hired by the Hazelwood School District during the two school years 1972–1973 and 1973–1974, 15 (or 3.7%) were black, whereas the proportion of black teachers in St. Louis County was 5.7%. One of the questions facing the Court was whether these two percentages were substantially different.

One way to begin analyzing such a question is to calculate the probability that a disparity this large or larger could occur. If one regards the 405 hires as a random sample from a pool of qualified teachers containing 5.7% blacks, then the probability of selecting 15 or fewer blacks may be calculated to be 4.6%. Such a probability may be calculated directly by using a *binomial probability distribution* (see Mosteller, Rourke, and Thomas, 1970, Chapter 4).

Such direct calculations are often tedious, and in practice the significance of such a disparity is often judged by an approximate procedure that involves computing a basic statistical quantity called the *standard deviation.* This measures how variable the data are. One then notes how many standard deviations the observed number of individuals (here 15) is from the expected number (here $405 \times 0.057 = 23$). A basic rule of thumb is that if individuals *are* selected at random, then 95% of the time the observed number will lie within 2 standard deviations of the expected number (more than 2 standard deviations above occurs about 2.5% of the time and more than 2 standard deviations below about 2.5%). Conversely, if the discrepancy exceeds 2 standard deviations, then a comparatively rare event has been observed and this calls into question the assumption that individuals were chosen at random; the difference is then said to be statistically significant at the 5% level.

The formula for the standard deviation in this setting is $\sqrt{np(1-p)}$, where n is the number of individuals, and p is the proportion in the population. For example, in *Hazelwood,* $n = 405$ and $p = 0.057$, so that the standard deviation in this case is $\sqrt{(405)(0.057)(0.943)}$ or 4.67. In a footnote (number 17), the Supreme Court noted that the difference between the number of blacks hired and the number expected ($23 - 15 = 8$) was less than 2 standard deviations, which would indicate that the difference is not statistically significant.

An important question that arose in *Hazelwood* was identifying the appropriate comparison population or relevant labor market. If one includes the city of St. Louis along with St. Louis County (as the government had argued), the percentage of black teachers increases from 5.7% to 15.5%. (This was in part because the city of St. Louis school system had attempted to maintain an

approximately 50% black staff.) Obviously in this case the disparity would achieve statistical significance. Because the disparity between 3.7% and the smaller labor market percentage of 5.7% was less than 2 standard deviations, the Court concluded that in this case the difference might be "sufficiently small to weaken the government's other proof, while the disparity between 3.7% and 15.5% may be sufficiently large to enforce it." Since this issue had never been considered by the lower courts, the appellate court decision was reversed, and the case was remanded to the district court for retrial.

Thus the value of statistical evidence as proof in an employment discrimination case can depend in a very sensitive way on which "relevant" labor market is chosen as a standard for comparison. In scientific sampling, where a random sample is drawn from an unambiguous and prespecified population, this standard of comparison must, of necessity, be clearly identified prior to drawing the sample. But in cases such as *Hazelwood,* what one begins with is the sample—the hiring rate for a protected class—and the population or relevant labor market is a hypothetical construct from which the hired employees are imagined to have been drawn. As the text of the *Hazelwood* decision makes clear, the choice of the relevant labor market to be used as a standard of comparison can be a complex question not necessarily admitting of a unique, simple, or unambiguous answer.

When the disparities observed in an employment discrimination case do not exceed those readily accounted for under random sampling, the result is some evidence against the presence of discrimination. If, on the other hand, a disparity is flagged as statistically significant, such a result needs to be interpreted with caution. For one thing, a hiring process need not be random in a statistical sense in order for it to be fair or unprejudiced. Fair processes of hiring can correspond to methods of sampling that will average to the correct expected value, but have larger standard deviations than those of simple random samples (see Meier, Sacks, and Zabell, 1986). For another, "statistical significance" merely means that there is *some* difference, but not that this difference necessarily has practical or legal significance. A difference of one or two percentage points in either direction, for example, in the pass rates of men and women on a physical dexterity test would scarcely be remarkable. As the appellate court in *United States* v. *Test* (550 F. 2d 577, 10th Cir. 1976) observed, "The mathematical conclusion that the disparity between these two figures is 'statistically significant' does not, however, require an *a priori* finding that these deviations are 'legally significant.' "

TWO-SAMPLE COMPARISONS OF APPLICANT DATA

Because deciding on the relevant labor market can involve delicate issues of judgment, the courts have expressed a strong preference for comparing applicant data with applicant data whenever possible (for example, what percentage of female applicants are hired versus the percentage of male applicants). This avoids having to decide on what the relevant comparison population is

and addresses a much more relevant question: Given the actual applicant pool, how did the protected class do? In the previous section we compared the percentage in one group with a well-determined population value, and thus dealt with a one-sample problem. Here we are comparing percentages of female and male applicants hired and so we are comparing the outcomes for two samples and we must allow for the variability from both sources. We discuss this shortly.

For example, in *Connecticut* v. *Teal* (457 U.S. 440, 1982)—another Supreme Court case—a written examination used to determine eligibility for promotion was challenged on the ground that it had a disparate impact on black examinees. Of 48 blacks who took the test, 26 passed (54.2%), while of 259 whites who took the test, 206 passed (79.5%). In this case the question is whether, if in the long run blacks and whites pass the test at the same rate, a sample difference this large or larger could occur. In *Teal,* the difference in pass rates is 79.5% − 54.2% = 25.3%, which is 3.76 standard deviations, well in excess of the "2 or 3 standard deviations" benchmark enunciated by the Supreme Court in its *Hazelwood* decision.

In many cases involving small businesses or short periods of time, the samples involved are too small to justify the use of the approximation that lies behind the standard deviation benchmark. Then recourse may be had to the *Fisher Exact Test* (see Mosteller, Rourke, and Thomas, 1970). The Fisher Exact Test views those individuals hired or promoted as a random sample from the applicant pool and calculates the probability that a disparity in hire or promotion rates at least as large as the one observed could arise.

Statistical analysis in cases involving small numbers can be particularly helpful because on many occasions intuition can be highly misleading. It is often thought that with small samples no useful information can be extracted, but it all depends on the size of the disparity involved. For example, in *Dendy* v. *Washington Hospital Center* (431 F. Supp. 873, D.D.C. 1977), the following table resulted:

	Selected	*Rejected*	*Pass Rate*
Blacks	4	5	44%
Whites	26	0	100%

The difference in pass rates for the two groups is obviously substantial, but because there are only 9 blacks in the entire sample, intuition might suggest that the difference could easily have happened by chance. In fact, such a disparity is extremely unlikely: In *Dendy,* the significance probability resulting from the Fisher Exact Test is .0004 (that is, there are no more than 4 chances in 10,000 of a difference this large or larger arising when treatment of the two groups is alike). (The history of *Dendy* is interesting: At the district level, the court—Judge John Sirica of Watergate fame presiding—rejected the statistical evidence on the ground that the sample involved was too small, but he was later overruled when the case was reviewed at the appellate level.)

Although consideration of applicant data avoids the problem of determining the relevant labor market, it can suffer from its own problems. Paradoxically, affirmative action programs aimed at encouraging minority groups to apply for positions can lead to applicant pools with differing levels of ability. In *Washington* v. *Davis* (426 U.S. 229, 1976), for example—another Supreme Court case—a test used by the District of Columbia Police Department was ruled to be nondiscriminatory, despite clear evidence that it had a disparate impact, partly because the pools may have had differing levels of ability.

SIMPSON'S PARADOX AND AGGREGATION

Large corporations and companies are almost always divided into divisions, subdivisions, departments, and so on. Except for high-level managerial positions, employment decisions usually take place at the departmental or divisional level. Analyzing aggregate employment data in such companies can give rise to a curious phenomenon known as *Simpson's paradox* (sometimes also referred to as *spurious correlation*).

An instructive and surprising example of Simpson's paradox occurred at the University of California at Berkeley in the 1970s. Examination of applicant data for a 1973 quarter revealed that the overall rate of admission for female applicants to the graduate school was substantially less than the rate of admission for male applicants (see Table 1).

Which departments at Berkeley were responsible for this imbalance? Surprisingly, admission data for individual departments showed that admission rates for males and females were comparable; indeed, in some departments admission rates were significantly *higher* for women!

This paradoxical situation actually has a very simple explanation. At Berkeley, women had applied more often to departments, such as English and History, with large numbers of applicants and correspondingly low rates of admission, while men had applied more often to departments, such as Mathematics and Physics, with fewer numbers of applicants and much higher rates of admission. A simple hypothetical example will illustrate the phenomenon:

	Mathematics			*English*			*Combined*		
	Admit	*Deny*	*%*	*Admit*	*Deny*	*%*	*Admit*	*Deny*	*%*
Males	90	10	90	1	9	10	91	19	83
Females	9	1	90	10	90	10	19	91	17

In this example, the admission rates for men and women are the same in each department: 90% in Mathematics, 10% in English. Nevertheless, because women applied more often to English and men more often to Mathematics, and because the overall admission rates for the two departments differed, the aggregate admission rates for men and women were substantially different.

Table 1 Applicants for graduate admission to the University of California, Berkeley, Fall 1973

Sex of Applicant	*Number of Applicants*			*Percentage*
	Admitted	*Denied*	*Total*	*Admitted*
Male	3,738	4,704	8,442	44%
Female	1,494	2,827	4,321	35%

Source: Bickel, Hammel, and O'Connell (1975).

Because of Simpson's paradox, it is appropriate that employment data in discrimination cases be analyzed at the level where decision-making actually takes place. Despite its appealing simplicity, considering only aggregate data can be highly misleading: fair employment practices can be made to appear unfair, and unfair employment practices can be disguised.

DISCUSSION

The use of statistics in employment discrimination cases reveals a surprising spectrum of issues involving application and interpretation, only some of which we have been able to explore here: the relevant comparison population, the appropriateness of the test of statistical significance based on random sampling, effects of aggregation, and controlling for relevant explanatory variables such as skill, to list but a few. The failure to take such problems into account can often result in a highly misleading picture, as in Simpson's paradox. Nevertheless, carefully used and appropriately interpreted, statistical methods provide some useful tools for the courts in their task of determining when prohibited discrimination in hiring, promotion, or firing has taken place.

PROBLEMS

1. In a recent case involving a suburban police department, a physical performance examination was challenged on the ground that it discriminated against women. While approximately 80% of 150 male applicants passed the exam during a six-year period, only 3 of 10 women passed during the same period. If the disparity is real, how could use of the examination be defended?
2. In the Berkeley graduate admissions case, could the different admission rates for departments be used as an argument for the presence of discrimination? Argue pro and con.
3. In 1968 Dr. Benjamin Spock and several associates were tried in Boston for conspiracy to violate the Military Service Act of 1967 (see *United States* v. *Spock,* 416 F. 2d 165, 1st Circuit 1969). Although more than 50% of adults in the Boston area were female, there were no women on Dr. Spock's jury!

This reduction came about in three stages: due to statutory disqualifications only 30% of the eligible jurors were female (stage 1); the jury list of 100—from which the jury was selected—contained only 9 women (stage 2); and—as a result of peremptory challenges—the final jury contained no women at all (stage 3). This was a matter of considerable concern to Dr. Spock's attorneys since women were thought to be favorably inclined toward their client, both because of his well-known books on child care and because polls indicated that women were more opposed to the Vietnam War than were men.

If you were an attorney for Dr. Spock, how would you attack the progressive diminution of women at each stage in the jury selection process? In particular, for stage 2, how large is the discrepancy in standard deviation units when a list of 100 jurors, drawn at random from a pool containing 30% women, contains 9 or fewer women?

4. In both employment and jury discrimination cases, the usual statistical calculations presuppose that the individuals in question (the persons hired or the jurors selected) are a random sample from the applicant or potential juror pools. In what ways other than discrimination might this assumption be violated in each case? At what stages in jury selection does the assumption seem more reasonable in jury discrimination cases?

5. From 1974 to 1978 the federal income tax rate for individuals decreased in every income category, yet the average overall rate of taxes actually collected increased from 14.1% to 15.2% (Wagner, 1982)! Explain why this is—in disguised form—an instance of Simpson's paradox.

REFERENCES

D. C. Baldus and J. W. L. Cole. 1980. *Statistical Proof of Discrimination.* Colorado Springs, Colo.: Shepard's. [The definitive reference.]

Peter J. Bickel, Eugene A. Hammel, and J. William O'Connell. 1975. "Sex Bias in Graduate Admissions: Data from Berkeley." *Science* 187:398–404. [Reprinted, with an interesting discussion between William Kruskal and Peter Bickel, in *Statistics and Public Policy,* W. B. Fairley and F. Mosteller, eds. 1977. Reading, Mass.: Addison-Wesley, pp. 113–130.]

Paul Meier, Jerome Sacks, and Sandy L. Zabell. 1986. "What Happened in *Hazelwood*: Employment Discrimination, Statistical Proof, and the 80% Rule." In *Statistics and Law,* M. DeGroot, S. Fienberg, and J. Kadane, eds. 1986. New York: Wiley, pp. 1–40. [Discusses some of the issues raised in this article in detail. The entire volume is a valuable source of information about recent legal uses of statistics.]

Frederick Mosteller, Robert E. K. Rourke, and George B. Thomas, Jr. 1970. *Probability with Statistical Applications,* 2nd ed. Reading, Mass.: Addison-Wesley. [An exceptionally lucid textbook with many examples.]

Clifford H. Wagner. 1982. "Simpson's Paradox in Real Life." *The American Statistician* 36:46–48. [Discusses several real-life instances of Simpson's paradox.]

Hans Zeisel. 1969. "Dr. Spock and the Case of the Vanishing Women Jurors." *The University of Chicago Law Review* 37:1–18. [A classic jury discrimination case.]