STATISTICAL METHODS FOR ANALYZING CLAIMS OF EMPLOYMENT DISCRIMINATION

JOSEPH L. GASTWIRTH*

This paper shows that in several recent EEO cases, lawyers and the courts have not used as powerful a statistical test of discrimination as they could have. The author describes two methods for combining into a single measure the results of statistical analysis of each of several data sets common in EEO cases, such as the hiring or promotion rates of minorities and majorities in each of several occupations in a company. He argues that these methods—combining one-sample binomial tests and the Mantel-Haenszel procedure—are more appropriate and usually more powerful than other tests of significance, such as Fisher's, that have been used in many EEO cases. He illustrates his argument with data from several of those cases.

SINCE 1977, when the U.S. Supreme Court used the binomial model to analyze data on the race of jurors in *Castenada v. Partida*[1] and the race of newly hired teachers in *Hazlewood,*[2] statistical testing to determine the significance of a difference between the observed and expected numbers of minorities or women hired or promoted has become routine in equal employment opportunity (EEO) cases. Indeed, recent court decisions have discussed the degree of statistical disparity that the plaintiff's statistical data must show in order to establish a prima facie case of discrimination, in cases of both disparate treatment and disparate impact.[3] According to the order of presentation of evidence the Supreme Court outlined in *McDonnell-Douglas v. Green,*[4] if plaintiffs establish a prima facie case, the defendant can rebut it by pointing out serious flaws in their data or statistical analysis or by submitting an alternative analysis that uses more accurate data on productivity characteristics. Those data should demonstrate that any remaining difference between the groups is not significant. Plaintiffs then are given the opportunity to refute the defendant's analysis.

*The author is Professor in the Departments of Statistics and Economics, George Washington University. This research was supported in part by National Science Foundation grants awarded to that university.

[1]*Castenada v. Partida,* 430 U.S. 482 (1977).

[2]*Hazlewood School District v. United States,* 97 S. Ct. 2736 (1977).

[3]The Supreme Court introduced these two classifications in *Teamsters v. United States,* 431 S. Ct. 431 at 335 n. 15 (1977), to distinguish cases involving an apparently neutral requirement that eliminates proportionately more minority members than majority members and is not justifiable by business necessity (disparate impact) from cases involving a general claim that the defendant purposely treated minorities less favorably than others (disparate treatment).

[4]*McDonnell Douglas Corp. v. Green,* 41 U.S. 792 (1983).

Industrial and Labor Relations Review, Vol. 38, No. 1 (October 1984).
0019-7939/84/3801 $01.00

The procedures involved in litigating employment cases can be more complicated than simply analyzing the minority composition of juries. Litigants may need to analyze hiring data for a number of occupations, each having its own minority availability;[5] promotion data over a range of salary or grade levels; or hiring data for several years during which the minority proportion of the qualified labor force or of actual applicants may have changed or the applications may have become stale. Sometimes the sample sizes for each subcategory (for example, persons hired during one year or in a particular occupation) are so small that meaningful statistical analyses for the individual categories are virtually impossible to make; yet at the same time the totality of data should be examined to ascertain whether it reflects a consistent pattern of minority underrepresentation.[6] When courts require plaintiffs to show a statistically significant disparity for each year of data or for each occupation separately, they are using a more stringent standard than the "two to three" standard deviations[7] mentioned in *Castenada* and *Hazelwood*. Combination tests may therefore aid plaintiffs in establishing a prima facie case of discrimination across a variety of jobs. On the other hand, an employer may be able to use a proper combination of tests of minority utilization in all relevant job categories to rebut evidence of minority underrepresentation in one of them.

This paper will describe and illustrate the use of two procedures for combining the binomial and hypergeometric[8] data typically employed in EEO cases. Since the results of both methods can be expressed as standard deviations, they are extensions of the approach the Court adopted in *Castenada* and *Hazelwood*. Moreover, both methods are more appropriate and usually more powerful[9] than Fisher's test,[10] which has been used in several EEO case. That test is based on the one-tailed significance levels or prob-values (the probability of data as extreme as the observed data occurring by chance, for example, the probability of

[5]The need to aggregate data from a variety of occupations also arose in assessing the impact of a consent decree. See Casey Ichniowski, "Have Angels Done More? The Steel Industry Consent Decree," *Industrial and Labor Relations Review*, Vol. 36, No. 2 (January 1983), pp. 182–98.

[6]Indeed, in *Vuyanich v. Republic National Bank*, 505 F. Supp. 224 at 376 (N.D. Tex. 1980), Judge Higgenbotham noted that "in many job families zero black hires fell within the randomness range so that statistical significance could never be present," and he therefore had to aggregate the data. The need to combine the results of statistical analyses of hiring in different occupations was noted by the Fourth Circuit in *EEOC v. American National Bank*, 652 F.2d 1176 at 1194 (4th Cir. 1981), and by Arthur Smith, Jr., and Thomas G. Abram, "Quantitative Analysis and Proof of Employment Discrimination," *University of Illinois Law Review*, No. 1 (1981), p. 33. Smith and Abram did not offer any suggestions of how to combine such data.

[7]The standard deviation of a distribution measures the spread of the population about its mean. For data from a normal distribution, observations that are 2 (3) or more standard deviations from the mean (expected value) occur about 5 (1) percent of the time. The 2- to 3-standard-deviation rule arose from the Court's use of the normal approximation to the binomial in *Castenada*. For further discussion, see David C. Baldus and James W. C. Cole, *Statistical Proof of Employment Discrimination* (New York: McGraw-Hill, 1980), pp. 294–97, and yearly supplements.

[8]In situations where selections are made from a relatively small and fixed pool of eligible candidates divided between two or more groups, the selection of a member of one group decreases the probability that a member of that group will be the next choice. In that case, the hypergeometric distribution gives the correct probability. When the pool of available persons is much larger than the number of selections, the binomial distribution yields a good approximation of the hypergeometric. When those selected form a sizable fraction of the total pool of eligible candidates, however, the binomial distribution is less accurate and understates the statistical significance of the data. For formulas and illustrations of the use of the hypergeometric distribution, see Elaine Shoben, "Differential Pass-Fail Rates in Employment Testing: Statistical Proof Under Title VII," *Harvard Law Review*, Vol. 91, No. 4 (February 1978), pp. 793–813; and Sidney Siegel, *Non-Parametric Statistics* (New York: McGraw-Hill, 1956), pp. 96–104.

[9]The power of a test (the probability of rejecting the null hypothesis when it is false, for example, finding minorities are underrepresented when their share of new hires is less than their proportion of the qualified labor force) has not been emphasized in legal decisions, although it underlies Judge Higgenbotham's observation in *Vuyanich* that if statistical significance cannot be obtained, calculating a statistical test is meaningless.

[10]Fisher's test is based on the probability that the product of the prob-values is as small as that calculated from the observed data. For details, see Robert Rosenthal, "Combining Results of Independent Studies," *Psychological Bulletin*, Vol. 85, No. 1 (January 1978), pp. 185–93.

few or fewer minorities being hired from the available labor pool) of the individual tests. Fisher's test also assumes that the underlying data come from a continuous distribution, such as the normal distribution, and it thus tends to be biased toward nonsignificance when used on the discrete or court data[11] that litigants employ in EEO cases.

Combining One-Sample Binomial Tests

When the minority fraction, p, of the relevant labor pool remains the same for the entire time period in question, pooling all years of data into one sample, as the Supreme Court did in *Castenada,* is the most powerful statistical technique that can be used.[12] When the availability fraction changes over time, however, or when the status of minority employees in several occupations, each with its own availability percentage of the external labor market, is at issue, the individual data sets cannot be treated as a large sample from the same binomial population. In such cases, we can obtain an overall picture by combining the results of the individual binomial tests. This is accomplished by comparing the sum of the differences between the actual and expected numbers of minority hires in each data set with its standard deviation. This method is thus a generalization of the usual approach.

Let n_i be the sample size (such as the number of new hires or employees) in the ith data set; let $i = 1, \ldots, k$, where k is the number of data sets being combined; let p_i be the minority fraction of the population sampled (p_i, for example, is the minority availability in the ith occupation); and let A_i be the number of minorities in the ith data set (for example, new hires). Each A_i has a binomial distribution with an expected value of $n_i p_i$ and a standard deviation of $\sqrt{n_i p_i(1 - p_i)}$, so their sum has a mean of $\Sigma n_i p_i$ and a standard deviation of $\sqrt{\Sigma n_i p_i(1 - p_i)}$, and it can be approximated by a normal distribution with the same mean and variance. Thus, we can compute the following normal variable (with continuity correction)[13] in standard-deviation units to make statistical inferences as the court did in *Castenada* and in *Hazelwood*:

$$(1) \qquad Z = \frac{\sum_{1}^{k} A_i - \sum_{1}^{k} n_i p_i + 1/2}{\left[\sum_{1}^{k} n_i p_i (1 - p_i)\right]^{1/2}}.$$

When all the p_is are the same p, the statistic given in Equation 1 reduces to the usual procedure for the pooled sample of size $N = \Sigma n_i$, from a binomial distribution with mean np and standard deviation $\sqrt{np(1 - p)}$. The next two sections illustrate the use of Equation 1 with data from two EEO cases.

Cooper v. University of Texas at Dallas

In this case,[14] the plaintiff charged the defendant university with sex discrimination in hiring faculty members and submitted data comparing the hires during 1976–77 with data on persons receiving doctorates in 1975 (from which the plaintiff derived availability proportions). Table 1 summarizes those data as reported in the opinion.

The plaintiff argued that those data established a prima facie case of discrimi-

[11]See W. Allen Wallis, "Compounding Probabilities from Independent Significance Tests," *Econometrica,* Vol. 10, Nos. 3–4 (July–October 1942), pp. 229–48; Egon S. Pearson, "On Questions Raised by the Combination of Tests Based on Discontinuous Distribution," *Biometrika,* Vol. 37, Nos. 3–4 (December 1950), pp. 383–98; and Henry O. Lancaster, "The Combination of Probabilities Arising from Data in Discrete Distribution," *Biometrika,* Vol. 36, Nos. 3–4 (December 1949), pp. 372–82.

[12]Jacobus Oosterhoff, "Combination of One-Sided Statistical Tests," No. 28 (Amsterdam: Mathematical Centre Tracts, 1968), p. 37.

[13]Since the binomial distribution is discrete, that is, it can only take on integer values and the normal curve used to approximate it is continuous, statisticians improve the approximation by considering each possible value of the binomial as being spread over an interval. For example, three hires are considered an interval (2.5, 3.5). For further discussion, see Frederick Mosteller, Robert E. K. Rourke, and George B. Thomas, Jr., *Probability with Statistical Application* (Reading, Mass.: Addison-Wesley, 1970), pp. 275–90.

[14]482 F. Supp. 187 (N.D. Tex. 1979), *aff'd,* 643 F.2d 1039 (5th Cir. 1981).

Table 1. Plaintiff's Hiring Data in *Cooper v. University of Texas at Dallas.*

University Division	*Availability* (p_i)	*Total Hires* (n_i)	*Actual Female Hires* (A_i)	*Expected Female Hires* (n_ip)	*Difference* $A_i - n_ip_i$	*Standard Deviation*	*Prob-value*
Arts-Humanities	.383	48	14	18.38	− 4.38	−1.30	.097
Human Development	.385	32	12	12.32	− .32	− .12	.452
Management-Administration	.043	26	0	1.12	− 1.12	−1.09	.319
Natural Science	.138	38	1	5.24	− 4.24	−2.13	.025
Social Science	.209	34	6	7.11	− 1.11	− .47	.319
Total		178	33	44.17	−11.17		

Source: 482 F. Supp. 187 (N.D. Tex, 1979), *aff'd*, 643 F.2d 1039 (5th Cir. 1981).

nation. Even though the data indicate a statistically significant disparity in only one division (Natural Science), the plaintiff argued that a prima facie case was established because the number of female hires was below its expected value in all divisions. The plaintiff concluded that the data were significant at the .032 level (one-tailed). Had the plaintiff focused instead on the total difference between actual and expected hires and used the generalized binomial test of Equation 1, she could have shown, by comparing the total difference (− 11.17) + .5 = − 10.67 with its standard deviation of 5.489 (calculated using Equation 1), that the difference between the actual and expected number of female hires was 1.94 standard deviations, corresponding to a two-tailed probability level of .052, which just misses significance at the .05 level. On the other hand, if the plaintiff had used Fisher's method to analyze the data in Table 1, she would have obtained a nonsignificant result corresponding to a two-tailed prob-value of about .10, confirming the criticisms of that test noted earlier.

Reporting the prob-value is more meaningful than simply stating that a test yields significance at the .05 (or .01) level. Indeed, the amount of rebuttal evidence required to refute a prima facie case is often suggested by the strength of the prob-value. In the case at hand, however, even the use of the generalized binomial test would not have been sufficient to carry the day for the plaintiff. In the end, the defendant successfully rebutted the plaintiff's charge by showing that her data included hires at senior ranks but measured availability proportions solely on data for recent doctoral recipients—a labor force comprising only those candidates available for the position of assistant professor.

Rivera v. City of Wichita Falls

The generalized binomial procedure could also have clarified the analysis in *Rivera,*[15] a case that raised the issue of the degree of statistical disparity required to establish a prima facie case. The Fifth Circuit analyzed minority hires in three job classifications by grouping them into two types, with availability percentages of Spanish-surnamed workers of 3 and 2 percent, respectively. The court tested the data in each category separately and found no significant minority underrepresentation. Table 2 reports the basic data in the opinion.

The normal approximation (Equation 1) yields a Z score of .97, or just under one standard deviation, which is not close to statistical significance at the usual .05 or .01 level. This conclusion is confirmed by an exact calculation yielding a one-sided prob-value of .15. Thus, the combination procedure places the court's findings of no prima facie case on firmer statistical

[15] 665 F.2d 531 (5th Cir. 1982).

Table 2. Hiring Data in *Rivera v. City of Wichita Falls.*

Job Class	*Availability*	*Total Hires*	*Actual Hispanic Hires*	*Expected Hispanic Hires*	*Difference*	*Standard Deviation*
1	.03	14	0	.42	.42	.64
2 and 3	.02	146	1	2.92	1.92	1.69
Total		160	1	3.34	2.34	

Source: 665 F.2d 531 (5th Cir. 1982).

grounds, by demonstrating that the disparity in the total number of relevant positions was not statistically significant.

In some cases, the filing of a charge or a suit may change the personnel practices of the defendant organizaton or the percentage of minority applicants, or both. In such instances, one should examine the data in light of these events and group the years appropriately.[16]

The Mantel-Haenszel Procedure

To analyze promotion, applicant flow (hiring), and discharge rates, courts have used the chi-square and Fisher-exact tests[17] to evaluate the significance of differences between data on majorities and minorities. In order to subdivide the employees or applicants into groups with similiar qualifications, it is necessary to combine the results of the separate tests of significance to reach a general conclusion. This section describes the Mantel-Haenszel procedure,[18] a summary test used to distinguish a systematic difference between rates derived from several 2 × 2 tables.

In addition to reporting the difference between the actual and expected number of minority appointments, which measures the number of positions the minority group lost by having a lower selection rate than the majority, the procedure yields an estimate of a relative measure of disparity:[19] the odds ratio, which compares the odds of a minority person being hired or promoted with the odds for a member of the majority group. This procedure can be illustrated with data from a case involving possible discrimination in promotions.

In *Agarwal v. McKee,*[20] Judge Orrick asserted that the plaintiffs established a prima facie case of promotion discrimination because the rate of promotions for minorities was significantly lower than that

[16]The Supreme Court stated in *United Airlines v. Evans,* 431 U.S. 553 (1977), that a "discriminatory act which is not made the basis for a timely charge is the legal equivalent of a discriminatory act which occurred before the statute was passed. It may constitute relevant background evidence in a proceeding in which the status of a current practice is at issue, *but separately considered,* it is merely an unfortunate event in history which has no present legal significance" (emphasis added). That statement has raised questions concerning the time period for which statistical data are relevant to the court's ultimate finding. In *Teamsters v. United States,* 431 U.S. 324 (1977), the court utilized data from July 2, 1965 until January 1, 1969 in its analysis of a suit filed in May 1968 because the challenged practices had gone unchanged until January 1969. If the data indicate that a change in policy occurred after the charge was filed, statistical evidence occurring after the policy change cannot be statistically considered as being from the same population as that existing before the change.

[17]See Shoben, "Differential Pass-Fail Rates."

[18]Nathan Mantel and William Haenszel, "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease," *Journal of the National Cancer Institute,* Vol. 22, No. 4 (1959), pp. 719–48. Their technique generalized previous results of William G. Cochran, "Some Methods for Strengthening the Common χ Square Tests," *Biometrics,* Vol. 10, No. 4 (December 1954), pp. 417–51, and it is an established method for combining the results of subgroup comparisons, as done in Baldus and Cole, *Statistical Proof.* For detailed references to the statistical literature, see Joseph Fleiss, *Statistical Methods for Rates and Proportions* (New York: John Wiley, 1973), and Brian Everitt, *The Analysis of Contingency Tables* (London: Chapman & Hall, 1977).

[19]For a discussion of both absolute and relative measures of treatment, see chapter 3 of Baldus and Cole, *Statistical Proof.*

[20]19 F.E.P. Cases 503 (N.D. Cal. 1977).

Table 3. Promotion Data From *Agarwal v. McKee* for the Period 1970–74.

		Minority			*Majority*		
Level	*Lowest Salary*	*Employed*	*Promoted*	*Rate*	*Employed*	*Promoted*	*Rate*
7	$21,360	19	3	15.8%	238	35	14.7%
8	19,704	39	7	17.9%	147	45	30.6%
9	17,628	87	17	19.5%	235	54	23.0%
10	15,660	143	34	23.8%	242	77	31.8%
Total		288	61	21.2%	862	211	24.5%

Source: 19 F.E.P. Cases 513 (N.D. Cal. 1977).

for nonminorities in the salary levels (7 through 10) in which three-fourths of the firm's professionals were employed. Although no formal statistical test was used in litigating this case itself, results from the Mantel-Haenszel (M-H) procedure corroborate the judge's opinion. Table 3 presents these results.

The M-H procedure computes the difference between the expected and the actual numbers of minority promotions for each salary level and the variance of that difference for each level. The sum of all the differences between actual and expected promotions corresponds to the numerator of the usual binomial procedure, but the M-H procedure also accounts for the difference in promotion rates by level. Because the variance of the sum of the differences is the sum of the individual variances, the standard deviation of the sum is the square root of the sum of the variances. The normal form of the M-H procedure divides the total difference between actual and expected promotions by its standard deviation, thereby yielding an analog of Equation 1. The difference between Equation 1 and the M-H procedure is that each of the random variables (the number of minority promotions in each level) being combined has a hypergeometric instead of a binomial distribution.

Table 4 presents the relevant formulas for a standard 2 × 2 table. Here, N is the total of $a + b + c + d$. Assuming both groups have the same promotion rate, that rate would be:

$$\frac{\textit{Number of Promotions}}{\textit{Total}} = \frac{a + c}{N}.$$

Thus, the expected number of minority promotions is:

$$(2) \qquad (a + c) \cdot (a + b)/N$$

and its variance is given by:

$$(3) \qquad [(a + c) \cdot (b + d) \cdot (c + d) \cdot (a + b)]/N^2(N - 1).$$

Table 4. Promotion Formulas for a Standard 2 × 2 Table.

	Promoted	*Not Promoted*	*Total*
Minority	a	b	$a + b$
Majority	c	d	$c + d$
Total	$a + c$	$b + d$	N

Table 5 presents data for the McKee employees at level 10. Employing those data in Equation 2 yields the expected number of minority promotions of (143) · (111)/385 = 41.23; the difference between actual and expected promotions for level-10 minority employees is therefore 34 − 41.23 = −7.23. This result means that level-10

Table 5. 2 × 2 Promotion Data for Level-10 Employees.

	Promoted	*Not Promoted*	*Total*
Minority	34	109	143
Majority	77	165	242
Total	111	274	385

Source: 19 F.E.P. Cases (N.D. Cal. 1977).

Table 6. Mantel-Haenszel Analysis of *Agarwal v. McKee* Promotion Data.

Level	*Expected Minority Promotions*	*Actual Minority Promotions*	*Difference*	*Variance*
7	2.81	3	+ .19	2.225
8	10.90	7	− 3.90	6.242
9	19.18	17	− 2.18	10.947
10	41.23	34	− 7.23	18.492
Total			− 13.12	37.906

Source: 19 F.E.P. Cases (N.D. Cal. 1977).

minorities received 7.23 fewer promotions than expected, assuming both groups had the same promotion rate. From Equation 3, the variance in the number of minority promotions equals $[(11) \cdot (274) \cdot (242) \cdot (143)]/[385^2 \cdot 384] = 18.49$.

Table 6 reports the Mantel-Haenszel calculations for the promotion data in Table 3. The normal form of the M-H test, with continuity correction, yields a value of:

$$(4) \quad \frac{\textit{Total Difference} + .5}{\sqrt{\textit{Variance}}} = \frac{-13.12 + .5}{37.906} = \frac{12.62}{6.156} = -2.05.$$

Hence, the difference (−13.12) between the actual and expected number is statistically significant at the .05 level, but not at the .01 level.

The data in Table 3 display two important characteristics that affect the M-H statistic. The *promotion rates vary by level* (usually declining as salary rises) and the distribution of minorities among the levels differs from that of nonminorities (proportionally more minorities are in the lower levels). If one ignored these aspects of the data and compared the overall promotion rates of 21.2 percent and 24.5 percent, one would not find the difference in the rates observed by Judge Orrick.

It should be emphasized that the defendant rebutted the charge of promotion discrimination by presenting data showing that minorities received their fair share of promotions within occupational categories; for example, they formed 29.5 percent of all professional and technical employees and received 32.5 percent of the promotions in that category. The plaintiffs, on the other hand, could have used the M-H procedure to test for equal promotion rates within each occupation, *by level*, to ascertain whether the occupational variable truly explained the difference in promotion rates.

EEOC v. Federal Reserve Bank

In a more recent case, *EEOC v. Federal Reserve Bank*,[21] the Fourth Circuit found that the plaintiff, the Equal Employment Opportunity Commission, failed to establish a prima facie case of discrimination because the Commission could not show a significant difference in promotion rates in each of two salary levels. Table 7 outlines the data from this case and the M-H analysis of promotions received during 1974–77 by persons employed at the beginning of each of those years.[22]

The normal form of the M-H test, Equation 4, yields separate minority shortfalls of −1.76 and −1.32 standard deviations in grades 4 and 5,[23] but combining the data for both grades shows that blacks received 12.13 fewer promotions than expected. That figure corresponds to a disparity of 2.19 standard deviations, which is equiva-

[21] 30 F.E.P. Cases 1137 (4th Cir. 1982).

[22] The opinion properly noted that the plaintiff's exhibit excluded persons who were promoted but later left the bank and included instead data on promotions given to all persons employed in the grade level as of the first of the year (see ibid., at 1152). Table 7 employs these data.

[23] Since those promoted were selected from a fixed pool of eligible candidates, the proper statistical model is the hypergeometric (based on sampling without replacement) rather than the binomial. The opinion may have misinterpreted the discussion in the 1982 supplement to Baldus and Cole, *Statistical Proof*, pp. 80–82, as implying that the binomial approximation to the hypergeometric was accurate in samples of 30 or more observations.

Table 7. M-H Analysis of *EEOC v. Federal Reserve Bank* Promotion Data.

Year	*All Employed*	*Minority Employed*	*Minority Promotions*	*Expected Minority Promotions*	*Difference*	*Variance*
			Grade 4			
1974	85	52	27	28.75	− 1.75	5.050
1975	51	31	8	8.51	− .51	2.470
1976	33	21	3	5.73	− 2.73	1.562
1977	199	73	1	2.00	− 1.00	.621
All Grade 4					− 5.99	9.703
			Grade 5			
1974	92	39	14	16.53	− 2.53	5.547
1975	107	53	14	13.87	+ .13	5.217
1976	79	41	19	19.20	− .20	4.974
1977	45	24	5	8.53	− 3.53	2.625
All Grade 5					− 6.14	18.362
Total					− 12.13	28.064

Source: 30 F.E.P. Cases 1152 (4th Cir. 1982).

lent to a two-tailed prob-value of just under 3 percent. In view of the fact that employees in grade 4 had to be promoted to grade 5 before they could be promoted again, differences in promotion rates at one level affected the sample size available for analysis at the next one and hence the possibility of detecting differences at the higher level. Restricting statistical analysis to individual levels obscures this aspect of the data. Moreover, combining the results of the yearly analyses into a summary M-H statistic is more reliable than pooling the four years of data into one sample because promotions in each year are taken from the pool of workers employed in that year; thus, variations in the minority fraction of the eligible pool resulting from employee turnover have little effect on the analysis.

Although the M-H procedure might therefore have helped the plaintiffs establish a prima facie case of discrimination, the court noted that even if they had established a statistically significant difference in promotion rates from grade 4, the court would have viewed that difference with caution because of the disproportionate number of blacks employed in cafeteria and service jobs rather than in clerical ones (in which there was greater opportunity for advancement). Unlike the decision in the *Agarwal v. McKee* case, the court did not require that the defendant provide specific data on promotions by occupational category. Again, application of the M-H procedure to the data in Table 7, if those data were further classified into clerical and nonclerical jobs, could ascertain whether the job category really reduced the statistical disparity to nonsignificance.

An important feature of the M-H test is that it allows us to test whether a *difference* exists in promotion, hiring, or layoff rates between groups after account has been taken of other factors, such as salary level, seniority, and education. When the circumstances of a case require that many potential influences be taken into account, the data in the individual 2 × 2 tables may become quite sparse, and models analogous to regression analysis may be required.[24]

[24]See, for example, Yvonne Bishop, Stephen Feinberg, and Paul Holland, *Discrete Multivariate Analysis* (Cambridge, Mass.: MIT Press, 1975); and David R. Cox, *Analysis of Binary Data* (London: Chapman & Hall, 1977).

Table 8. Promotion Data, by Race, from *Hogan v. Pierce.*

	Whites		*Blacks*	
Date of Promotion	*Eligible*	*Promoted*	*Eligible*	*Promoted*
July 1974	20	4	7	0
August 1974	17	4	7	0
September 1974	15	2	8	0
April 1975	18	1	8	0
May 1975	18	1	8	0
October 1975	30	1	10	0
November 1975	31	2	10	0
February 1976	31	1	10	0
March 1976	31	1	10	0
November 1977	34	1	13	0

Source: Plantiff's exhibit on file with D.C. District Court.

Hogan v. Pierce

The M-H test can also be used to test for a difference in promotion rates over time, even though the pool of eligible candidates may change during the period in question.[25] In *Hogan v. Pierce,*[26] the plaintiff established a prima facie case, in part by analyzing promotions from grade 13 to grade 14 from March 1972, the effective date of the amendments extending the Civil Rights Act to the government sector, through the date of the administrative complaint in 1977. The plaintiff established his case in part by enumerating the employees who possessed at least the minimum qualifications for each promotion at the time it was made.[27] The eligible candidates in Table 8 were employed in computer occupations and had at least one year's experience in the previous grade (13).

The plaintiff then applied the M-H test to the data in Table 8, obtaining:

$$\frac{\textit{Observed} - \textit{Expected} + .5}{\textit{Standard Deviation}} = \frac{4.517}{1.837} = 2.46.$$

Thus, if black and white eligible candidates had the same chance of promotion, the probability that no blacks would be promoted is about .007, a statistically significant result.[28] In his opinion accepting the plaintiffs' analysis, Judge Robinson also noted that the analysis demonstrated that the pool of eligibles was large enough that "zero" promotions could not have occurred by chance, even though the total number of promotions over the relevant time period was relatively small (18).

The Odds Ratio

Before concluding, it is important to discuss the "odds ratio," a useful interpretation of the difference between two rates. First, recall that if an event has probability p of occurring, then the probability it will not occur is $1 - p$, and the odds of the event's occurring are p to $1 - p$. In betting language, this means that when you bet on the occurrence of the event, the ratio of your bet to your potential winnings should

[25] Thus, one can combine the results of every promotion decision and avoid the fragmentation issue raised by Judge Greene in *Trout v. Hidalgo,* 517 F. Supp. 8738 at n.35 (D.C. Cir. 1980), as well as the technical problems faced by the Court in *EEOC v. Federal Reserve Bank.*

[26] *Hogan v. Pierce,* 31 F.E.P. Cases 115 (D.C.D.C. 1983).

[27] This is the criterion established in *Davis v. Califano,* 613 F.2d 957 (D.C. Cir. 1979) and *Valentino v. United States Postal Service,* 674 F.2d 56 (D.C. Cir. 1982).

[28] The corresponding two-tailed test would have a prob-value of .014. Nonetheless, other researchers who examined similar data show that the M-H procedure with continuity correction tends to overestimate the prob-value, and they recommend omitting the correction. Their method yields a disparity of 2.73 standard deviations, which is significant at the .01 level (two-tailed test). See Lloyd Lininger, Mitchell Gail, Sylvan Green, and David Byar, "Comparison of Four Tests for Equality of Survival Curves in the Presence of Stratification and Censoring," *Biometrika,* Vol. 66, No. 3 (December 1979), pp. 419–28.

be $p/(1 - p)$. For instance, if $p = 1/3$, then $1 - p = 2/3$ and $p/(1 - p) = 1/2$; or the odds are 1 to 2 against the event's happening. Thus, a fair bet would offer you \$2 for every \$1 you wager on the occurrence of the event. This agrees with our intuition, since the chance that the event that will not take place is twice as large as the chance that it will.

If we consider two different rates (p_1 and p_2) of, for example, the promotion of majority and minority workers, respectively, the odds facing a majority member are $p_1/(1 - p_1)$, while those facing a minority member are $p_2/(1 - p_2)$. The ratio of these two odds,

$$(5) \quad \frac{p_2}{(1 - p_2)} \cdot \frac{(1 - p_1)}{p_1} = \frac{p_2}{p_1} \cdot \frac{(1 - p_1)}{(1 - p_2)},$$

is called the "odds ratio." When the odds ratio is less than one, minorities have less chance for promotion than majority workers. The reverse is true if the odds ratio is greater than one.

It should be noted that the odds ratio is related to the selection ratio p_2/p_1, which has been used to define the four-fifths rule.[29] The former ratio is *symmetric*,[30] however, since the odds ratio based on failure rates $(1 - p)$ is simply the reciprocal of the odds ratio based on the pass rates.

For data reported in a 2 × 2 table, as in Tables 4 and 5, p_2 can be estimated by $a/(a + b)$; $1 - p_2$ by $b/(a + b)$; p_1 by $c/(c + d)$; and $1 - p_1$ by $d/(c + d)$. Subsituting these estimates into Equation 5 shows that the odds ratio is estimated by:

$$(6) \quad \frac{a\,d}{b\,c}.$$

For the data in Table 5, the estimated odds ratio is $(34 \cdot 165)/(77 \cdot 109) = .668$, indicating that the odds of a minority employee in level 10 receiving a promotion are about two-thirds those of a majority employee.

In order to combine (1) the results of 2 × 2 tables indexed by i with N_i observations in the ith table with (2) an overall estimate of the odds ratio, the M-H procedure provides the estimate:

$$(7) \quad \frac{\sum_{i=1}^{k}(a_i d_i)N_i^{-1}}{\sum_{i=1}^{k}(b_i c_i)N_i^{-1}} = \sum_{i=1}^{k} w_i(a_i d_i/b_i d_i),$$

where

$$w_i = \frac{[(b_i c_i)N_i^{-1}]}{\sum_{i=1}^{k}(b_i c_i)N_i^{-1}}$$

is the weight given to the odds ratio from the ith table. Applying Equation 7 to the four-component 2 × 2 tables (one for each level) in Table 3 yields an overall odds ratio of .698. Thus, the odds of a minority employee at McKee receiving a promotion were only about 70 percent as favorable as those for majority employees. This would be a meaningful difference if an 80 percent criterion[31] (analogous to the four-fifths rule used for selection rates) were adopted.

The estimated odds ratio can be used to compare relative disparities across cases and is a useful supplement to formal statistical testing. As an example, consider the appellate decision in *EEOC v. American National*

[29]The federal "Uniform Guidelines on Employee Selection Procedures," 13 Fed. Reg. 38, 295–38, 309 (1978), indicate that if the selection rate of minorities is less than four-fifths that of the majority, adverse impact exists. The guidelines have received much comment: for example, Shoben, "Differential Pass-Fail Rates," and Richard D. Arvey, *Fairness in Selecting Employees* (Reading, Mass.: Addison-Wesley, 1979).

[30]This lack of symmetry of the selection ration p_2/p_1 was noted by Shoben, "Differential Pass-Fail Rates," and by Smith and Abram, "Quantitative Analysis," p. 33, note 129. When p_1 and $1 - p_1$ as well as p_2 and $1 - p_2$ are interchanged, the odds ratio using pass rates is a, and the ratio using failure rates is $1/a$. Thus, saying that members of group A have twice the odds of passing as members of group B is equivalent to saying that they have one-half the odds of failing.

[31]Unfortunately, there is no simple formula for translating selection ratios to odds ratios. For example, minority and majority selection rates of 20 and 25 percent translate to an odds ratio of .75, whereas the apparently corresponding rates of 40 and 50 percent yield an odds ratio of .667.

Bank.[32] The original panel found hiring discrimination in clerical positions, but a subsequent evenly divided (4–4) *en banc* opinion[33] denied the defendant a rehearing. The first panel found that the EEOC had laid out a prima facie case by comparing the minority fraction of clerical employees with their availability in the Suffolk and Portsmouth, Virginia areas. That panel rejected the bank's analysis of applicant-flow data, because data for only one year out of seven were available for Portsmouth and because the black fraction of hires was less than half their fraction of applicants in both areas.

In his dissent from the *en banc* decision upholding the finding of discrimination, Judge Widener asserted that the applicant flow should be analyzed.[34] He pooled all the available data on the Suffolk area into one sample and found a minority disparity of 3.7 jobs, or 1.45 standard deviations. Similarly, the judge analyzed the single year of available Portsmouth data and obtained a minority disparity of 4.3 jobs, equivalent to 2.16 standard deviations, and concluded that a prima facie case might have been shown in Portsmouth.

Had the M-H analysis for the combined data been used (Equation 4), a disparity of eight jobs and 2.38 standard deviations would have resulted. Moreover, Equation 7 calculates an odds ratio of .34, which is much smaller than those in other cases studied here (that is, versus .70 in *Agarwal* and .66 in *EEOC v. Federal Reserve Bank).* Thus, this analysis supports the finding of the original panel. It is interesting to note that although the eight positions minorities "lost" in *American National Bank* were fewer than the 13 in *Federal Reserve,* the M-H procedure yielded a more significant result in the former case—a result brought to light by comparing the odds ratios.

Conclusion

The procedures discussed here can assist courts in evaluating data in employment discrimination cases. They allow comparison of sets of related data by combining the results of statistical calculations of each data set without having to assume inaccurately, for example, that promotion rates are the same at all levels of a hierarchical system or discharge rates are the same for all employees regardless of seniority. The application of the procedures to actual case data demonstrates the importance of reporting the exact prob-value of the data, not just whether it was statistically significant at the .05 level (or differed by two standard deviations), as well as the need for care in using the normal approximation (standard-deviation analysis) for binomial data and the binomial approximation to the hypergeometric model.

By using these methods, the courts can also evaluate whether a party's explanation of a significant disparity in fact reduces that remaining disparity to statistical insignificance. The extra factor required by these methods will also require that more subgroups be analyzed,[35] but the results of the appropriate calculations can be combined into an overall test of whether any remaining difference between the minority and majority groups is statistically significant.

These combination methods, which are based on a measure of absolute disparity, also have an associated measure of relative

[32] 652 F.2d 1176 (4th Cir. 1981).

[33] 680 F.2d 965 (4th Cir. 1982).

[34] This kind of availability comparison was questioned in Joseph L. Gastwirth, "Estimating the Demographic Mix of the Available Labor Force," *Monthly Labor Review,* Vol. 104, No. 4 (April 1981), pp. 50–57, in part because pre–Civil Rights Act hires were properly removed from data on the firm's employees but the census statistics were not similarly adjusted. Thus, the availability of minorities and women for entry-level jobs was underestimated. For further discussion of the pitfalls in using unadjusted census data for determining availability, see Baldus and Cole, *Statistical Proof.*

[35] Since introducing more factors decreases the statistical power of the combined test (see, for example, Lininger et al., "Comparison of Four Tests"), courts should consider the relevance and potential additional explanatory power a new factor might have before rejecting an otherwise valid analysis showing a very large or small disparity.

disparity, the odds ratio.[36] This ratio should aid courts in assessing the meaningfulness of a disparity, as well as in comparing statistical conclusions based on data sets of varying sample sizes from similar cases.[37]

[36]The analog of Equation 7 yielding a summary odds ratio from the combination of binomial data sets is discussed in a forthcoming technical report by the author and Samuel W. Greenhouse (Washington, D.C.: Department of Statistics, George Washington University).

[37]Combination procedures, like all statistical methods, need to be used carefully and properly, not routinely. Statistically significant differences in one category should not be combined with unrelated tests either to widen the areas of alleged discrimination or to obscure a discriminatory policy. The court allowed commission of this second error in *Adams v. Gaudet,* 515 F. Supp. at 1138–39 and 1145 (W.D. La. 1981), where tests on promotion data were combined with tests on hiring data, thereby obscuring a statistically significant disparity in minority hiring in the three jobs at issue.