A CASE OF UNEXAMINED ASSUMPTIONS: THE USE AND MISUSE OF THE STATISTICAL ANALYSIS OF *CASTANEDA/HAZELWOOD* IN DISCRIMINATION LITIGATION

THOMAS J. SUGRUE AND WILLIAM B. FAIRLEY*

In *Castaneda v. Partida*[1] the Supreme Court of the United States for the first time used formal statistical methods to determine whether data showing an underrepresentation of a particular ethnic group among persons selected to serve on grand juries supported an inference of discrimination against that group in violation of the equal protection clause of the fourteenth amendment to the Constitution. Later the same term, the Court applied the particular statistical test used in *Castaneda* to issues of racial discrimination in employment in *Hazelwood School District v. United States*,[2] a case brought under Title VII of the Civil Rights Act of 1964.[3] The analytic approach adopted in the *Castaneda* and *Hazelwood* cases consists of the use of a particular statistical model to calculate both the "expected" results of an ethnically or racially neutral system and the likelihood of seeing differences from those expected results as large or larger than those actually observed. If the observed differences, considered in relation to the size of the sample to which they refer, are of such a magnitude

---

* Thomas J. Sugrue, B.S., Boston College, 1968, M.P.P., Harvard, 1975, J.D., Harvard, 1975, is a member of the District of Columbia Bar and practices at the firm of Wilmer, Cutler & Pickering. William B. Fairley, B.A., Swarthmore, 1960, Ph.D., Harvard, 1968, is an economist and statistician at Analysis and Inference, Inc., a research and consulting company in Boston, Massachusetts.

The authors acknowledge with appreciation the helpful comments on drafts of this article provided by Arlene Ash, David Baldus, William Lake, Jay Lapin, Michael Meyer, and Stephan Michelson. The authors alone, however, are responsible for any errors or omissions in the article.

[1] 430 U.S. 482 (1977).

[2] 433 U.S. 299 (1977).

[3] 42 U.S.C. §§ 2000e to 2000e-17 (1976). Title VII provides, in § 703(a)(1), that it is unlawful employment practice for an employer to "fail or refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin." 42 U.S.C. § 2000e-2(a)(1) (1976).

that they would be very unlikely to occur as the chance result of a neutral system,[4] the plaintiff is deemed to have made out a *prima facie* case,[5] thereby shifting to the defendant the burden of rebutting the plaintiff's statistical proof by demonstrating that it is "inaccurate or insignificant"[6] or by identifying nondiscriminatory factors that account for the observed differences.[7] The advantage of using a formal statistical model in such a case is that it allows the factfinder to draw inferences from the data that are more accurate and mathematically meaningful than those obtained through "eyeballing" the data or simple arithmetic comparisons.

Not surprisingly, in the wake of *Castaneda* and *Hazelwood*, the statistical model developed in those cases has been applied by lower courts and litigants to statistical evidence of discrimination in a wide variety of contexts.[8] The validity of the *Castaneda/Hazelwood* model, however, depends on a number of assumptions not spelled out by the Court that are not met in a large number of

---

[4] Such differences are said, as a matter of definition within the field of statistics, to be "statistically significant." An analysis of statistical significance of a difference is most helpful when it reports the probability that, if the difference were the result of a neutral system that operated in accordance with the assumptions of the statistical model chosen to describe it, a difference as large or larger than it would be observed. This probability is called the "descriptive significance level" associated with the difference as interpreted by the model. Sometimes the probability itself is not reported and instead only the result of comparing it to some chosen standard "level of significance," often taken to be 1%, 5% or 10%. If the probability is lower than the standard, then the difference is declared to be "statistically significant" at that level; if higher, it is declared to be "not statistically significant" at that level. Basic statistics texts which discuss the concept of statistical significance are F. MOSTELLER, R. ROURKE, & G. THOMAS, PROBABILITY WITH STATISTICAL APPLICATIONS (2d ed. 1970) (hereinafter cited as MOSTELLER, ROURKE & THOMAS) and G. SNEDECOR & W. COCHRAN, STATISTICAL METHODS (7th ed. 1980) (hereinafter cited as SNEDECOR & COCHRAN).

Moreover, in D. BALDUS & J. COLE, STATISTICAL PROOF OF DISCRIMINATION (1980) (hereinafter cited as BALDUS & COLE) the significance concept is discussed with specific application to discrimination cases. *See id.* at §§ 9.2-9.4. Statistical significance is a subtle and technical concept that when properly applied can be a powerful tool in appraising the evidentiary value of statistical differences. It must, however, be carefully distinguished from certain everyday meanings attributed to the term "significance," and it must be carefully applied to data in discrimination litigation, in most cases with the help of experts who are qualified in statistical analysis and preferably in the application of such analysis in a legal setting.

[5] Statistical proof alone may be sufficient to make out a *prima facie* case of discrimination if the disparities are sufficiently incriminating. Hazelwood School Dist. v. United States, 433 U.S. at 307-08. McKenzie v. Sawyer, 684 F.2d 62, 71 (D.C. Cir. 1982); Pouncy v. Prudential Ins. Co. of Am., 668 F.2d 795, 802 (5th Cir. 1982). Furthermore, statistical proof may be combined with other forms of proof, and plaintiffs often introduce nonstatistical evidence in order to buttress their statistical showing of discrimination and bring the "cold numbers convincingly to life." International Bhd. of Teamsters v. United States, 431 U.S. 324, 339 (1977).

[6] International Bhd. of Teamsters v. United States, 431 U.S. 324, 360 (1977); *see also* Dothard v. Rawlinson, 433 U.S. 321, 338-39 (Rehnquist J., concurring) (Employers "may endeavor to impeach the reliability of statistical evidence, they may offer rebutting evidence, or they may disparage in arguments or in briefs the probative weight which the plaintiffs' evidence should be accorded." *Id.*).

[7] International Bhd. of Teamsters v. United States, 431 U.S. at 360-61 n.46; Castaneda v. Partida, 430 U.S. at 497-99.

[8] *See* cases cited *infra* note 50.

"selection" situations in which the model appears at first blush to be applicable. This has led some courts and litigants to use this model incorrectly and to reach erroneous conclusions about the meaning of statistical evidence in particular cases.

This article analyzes the strengths and limitations of the statistical model applied in *Castaneda* and *Hazelwood*, reviews its use and misuse in discrimination litigation as a means of testing whether a selection system adversely affects particular groups, and suggests alternative models for those situations in which the *Castaneda/Hazelwood* model should not be applied.[9]

I. THE *CASTANEDA/HAZELWOOD* ANALYSIS

In *Castaneda*, the petitioner alleged that Mexican-Americans had been intentionally discriminated against in selection for grand jury duty in Hidalgo County, Texas.[10] Part of the petitioner's proof involved a showing that the number of Mexican-Americans selected as grand jurors was less than proportional to the representation of Mexican-Americans in the eligible juror pool.[11] A statistical test was employed to establish that the difference was large enough to rule out chance as a reasonable explanation.[12] Similarly, in *Hazelwood*, the plaintiffs alleged that blacks had been discriminated against in the hiring of teachers in a school district in Missouri.[13] Part of the proof in that case involved

---

[9] The specific questions to which such statistical models help provide answers are whether and to what degree observed disparities in the results of a selection system, which suggest discrimination, could be attributable to one or more chance mechanisms that are neutral with respect to the particular groups being compared rather than to systematic or purposeful non-neutral mechanisms. This article analyzes the reliability of the statistical model applied in *Castaneda* and *Hazelwood* for that purpose. It does not address how the answers to those questions should be combined with a consideration of the magnitude of the differences in selection results between groups (as evidenced by the size of the observed disparities) or other evidence of a non-statistical nature in the case to reach a finding on the ultimate issue of discrimination.

[10] Castaneda v. Partida, 430 U.S. at 483-84. The petitioner in *Castaneda* was a Mexican-American who had been convicted of a crime in a Texas state court. *Id.* at 485. He challenged his conviction — first in the state courts and then, after exhausting his state remedies, in the federal courts on a habeas corpus petition — on the ground that the grand jury that indicted him had been selected in a discriminatory fashion. *Id.* at 485-92. Prior to *Castaneda*, the Supreme Court had held that it is a denial of the equal protection of the laws guaranteed by the fourteenth amendment to the United States Constitution to try a defendant who has been indicted by a grand jury from which persons of his race or of an identifiable group to which he belongs have been purposefully excluded. *Id.* at 492-95 and cases cited therein.

[11] *Id.* at 485-89.

[12] *Id.* at 494-97.

[13] Hazelwood School Dist. v. United States, 433 U.S. at 301. *Hazelwood* was brought under Title VII of the Civil Rights Act of 1964. *See supra* note 3 and accompanying text. The Supreme Court has held that there are two different types of actionable violations of Title VII — disparate treatment and disparate impact. International Bhd. of Teamsters v. United States, 431 U.S. 324, 335 n.15. Disparate treatment occurs when an employer treats some people less favorably than others because of their race, color, religion, sex, or national origin. *Id.* Proof of discriminatory intent is required to establish a claim of disparate treatment. *Id.* Disparate impact, on the other hand, occurs when an employer adopts practices that, while neutral on their face, affect one group more adversely than others and cannot be justified by business necessity.

a showing that the number of blacks hired was less than proportional to the representation of blacks among qualified public school teachers in the relevant labor market.[14] The *Hazelwood* Court employed the same statistical test used in *Castaneda* to determine whether the difference was large enough to rule out chance as an explanation.[15]

It will be useful to discuss the *Castaneda* test in more detail. Over an eleven year period, the number of persons summoned to serve on grand juries in Hidalgo County was 870.[16] The Mexican-American percentage of the population of the county was estimated to be 79.1%.[17] The Court reasoned that if selections were made on a random basis from that population, about 79.1%, or 688 of the 870 grand jurors would be Mexican-Americans.[18] The actual number of Mexican-Americans called was 339.[19]

The Court in Castaneda found it useful to inquire about the probability of selecting so few Mexican-Americans under a hypothetical assumption that the 870 jurors were drawn randomly.[20] Since jurors were not in fact drawn randomly, but rather, under the Texas "key man" system, on the personal judgment of jury commissioners, random drawings were used by the Court as a standard or benchmark for a selection process that is free of discrimination.[21] The Court implicitly adopted a particular set of properties (or "model") for such random drawings: (1) a fixed probability of selection of a Mexican-American on each drawing (in this case 79.1%); (2) only two possible outcomes for each drawing, in this case Mexican-American and non-Mexican-American; and (3) an independent drawing each time, so that the chance of selection of a Mexican-American on each drawing is not affected by the outcomes of the prior drawings.[22] Under these assumptions, which define a "binomial model"

Id. Discriminatory intent is not a necessary element of a disparate impact claim. *Id.* Under either theory, statistical proof will normally play an important role in establishing or rebutting a claim of classwide discrimination. *Id.* *Hazelwood* was a disparate treatment case, *i.e.*, the plaintiffs there alleged that the defendant school district had intentionally limited its hiring of black teachers on the basis of race. *Id.* at 301.

[14] *Id.* at 303.

[15] *Id.* at 311 n.17. The Court in *Hazelwood* did not reach the question whether the defendants had actually engaged in discrimination but remanded the case to the district court for a determination, *inter alia*, of the geographic scope of the appropriate labor market. *Id.* at 313. The labor market proposed by the plaintiffs (which included the City of St. Louis) contained a much higher percentage of black teachers than did the labor market proposed by the defendants (which excluded the City of St. Louis). *Id.* at 310-11. The Court applied the statistical test employed in *Castaneda* to both these proposed labor markets in order to "highlight the importance of the choice" between the two. 433 U.S. at 311 n.17. The Court stated that its analysis demonstrated that if the former were adopted the statistical proof would support the Government's case that Hazelwood had engaged in discrimination, while if the latter were adopted the statistics would not support and might even weaken the Government's case. *Id.*

[16] Castaneda v. Partida, 430 U.S. 482, 487 n.7.

[17] *Id.* at 486.

[18] *Id.* at 495-96.

[19] *Id.* at 487 n.7.

[20] *Id.* at 496 n.17.

[21] *Id.* at 484.

[22] *Id.* at 496 n.17.

for drawings, the probabilities of drawing different numbers of Mexican-Americans are given by the "binomial distribution."[23]

In order to assess the possibility that the shortfall in the number of Mexican-Americans called to serve on grand juries could have occurred by chance, the Court used the binomial model to calculate a measure by which to calibrate the significance of the difference between the 339 observed and the 688 expected Mexican-American grand jurors.[24] This measure, the standard deviation, was calculated to be 12.[25] The difference between 688 and 339 was therefore 29 standard deviations, and the Court noted that "a detailed calculation reveals that the likelihood that such a substantial departure from the expected value would occur by chance is less than 1 in 10^{140}."[26]

The Court also noted a rule of thumb for the number of standard deviations that would imply a high enough improbability of the observed result to warrant rejection of the original hypothesis: "As a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the jury drawing was random would be suspect to a social scientist."[27]

II. The Binomial Model

The binomial model employed by the Court in *Castaneda* is one of the simplest and most useful analytic tools available to evaluate statistical evidence in discrimination cases.[28] The model can be applied to any situation that consists of a series of events, such as drawings or selections (which statisticians sometimes refer to as "trials") with the above-specified properties of two possible outcomes for each trial, fixed probabilities associated with each outcome, and independence among the trials. Such a series of events is called a "binomial experiment." Simple physical prototypes of binomial experiments include flipping a coin a number of times and drawing poker chips of two types — *e.g.*, of two different colors — from a fishbowl containing a large number of such chips. In the former case the two possible outcomes for each trial are heads or tails and the probability associated with each, assuming we are dealing with a fair coin fairly tossed, is .5. In the second case the two possible outcomes correspond to the two different colors of the chips, say black and white, and the probability associated with each outcome is given by the ratio of the number of chips of that color in the bowl to the total number of chips in the bowl. For ex-

---

[23] *Id.* The binomial model also requires that there be a fixed — as opposed to a randomly determined — number of drawings, *see* MOSTELLER, ROURKE & THOMAS, *supra* note 4, at 132. This condition has been of less interest in discrimination litigation and we do not treat it in this article.

[24] Castaneda v. Partida, 430 U.S. 482, 496 n.17.

[25] *Id.*

[26] *Id.*

[27] *Id.*

[28] For a thorough discussion of binomial models and distributions, *see, e.g.*, MOSTELLER, ROURKE & THOMAS, *supra* note 4, at 130-43, 270-91.

ample, if the bowl contains 10,000 chips, 6,000 of which are white and 4,000 of which are black, the probability of drawing a white chip — assuming the chips are otherwise identical and have been well mixed together in the bowl — would be .6 and the probability of drawing a black chip would be .4.[29]

For any given binomial experiment, once the number of trials and the probability on each trial of observing one of the two possible outcomes — let us call that outcome a "success"[30] — are specified, we can calculate an "expected" number of such successes by multiplying that probability by the number of trials.[31] For example, if we were to draw 10 chips from the bowl described above, the expected number of white chips would be 6, obtained by multiplying .6 by 10. And this is consistent with common sense — if 60 percent of the chips in the bowl are white, we would expect that *on average* 60 percent of the chips drawn from the bowl will be white. We would not expect, however, that every time we draw 10 chips from the bowl exactly 6 of them will be white. It is clear that any result from zero to 10 white chips is possible, although we would not expect all such possible results to occur with equal frequency. For example, we would expect that 5 or 7 white chips would be drawn fairly often, 4 or 8 white chips somewhat less frequently, and more extreme outcomes only on occasion. This commonsense insight into the variability of the results of drawing chips from a bowl corresponds to what statistical theory tells us about the relationship between the expected value of a binomial experiment and the actual observed outcomes. The results of a binomial experiment are, as the Supreme Court noted in *Castaneda*, "likely to fall in the vicinity of the expected value,"[32] and fall in such a way that the further away a result is from the expected value, the less frequently it will actually be observed as the outcome of

---

[29] The sum of the probabilities of the two outcomes of a binomial experiment will allways equal 1.00. This is so because a probability of 1.00 corresponds to a certainty, that is that an event will happen 100% of the time. Since for each trial in a binomial experiment there are only two possible outcomes, their joint probability — that is the probability one or the other outcome will occur — must equal 1.00. For example, in drawing chips from a bowl containing only white and black chips, it is certain that for each binomial trial (*i.e.*, for each drawing) the result will be a chip of one of these two colors. Furthermore, since the two outcomes are mutually exclusive, their joint probability is derived by simply adding their individual probabilities. *See* MOSTELLER, ROURKE & THOMAS, *supra* note 4, at 77-81.

[30] For convenience of reference, social scientists often refer to one binomial outcome as a "success" and the other as "failure." *See* MOSTELLER, ROURKE & THOMAS, *supra* note 4, at 130-31. These designations are arbitrary and have no effect on the binomial calculations, which are not affected by what labels one puts on the outcomes — *i.e.*, the probability of observing particular combinations of the two outcomes are determined in the same fashion regardless of which outcome is designated a "success" and which a "failure." But since it does make it easier to describe the calculations, we will at times use the success/failure terminology.

[31] The "expected value" of a statistical experiment, binomial or otherwise, refers to the long-run average result. That is, if the experiment were repeated a large number of times, we would expect that the average of the results would be close to and, as the experiment was repeated over and over, be converging on the expected value. Of course, as discussed in the text, the actual results of any single experiment will in most cases be somewhat lower or higher than the expected value.

[32] 430 U.S. at 496 n.17.

such an experiment. A measure of the extent of this "spread" of the distribution of possible results around the expected result -- that is, the extent to which the actual outcomes are likely to diverge from the expected value — is the distribution's standard deviation. As the Supreme Court noted in *Castaneda*, the number of standard deviations any particular result is from the expected result is inversely related to the probability of such an occurrence.[33]

The binomial model permits the calculation of an exact probability for each possible outcome of a binomial experiment, and the set of outcomes paired with probabilities defines that experiment's probability distribution.[34] For example, in the experiment discussed above of drawing 10 chips from a bowl, the probability of drawing exactly 6 white chips can be precisely determined as .251 — that is, if we repeated the experiment a very large number of times, we would expect 25.1% of those times exactly 6 white chips (and 4 black chips) would be drawn. Similar exact probabilities may be determined for each possible outcome from zero to 10 white chips.[35] Other than when the number of trials is fairly small, however, the computations required to calculate exact binomial probabilities are quite burdensome. Fortunately, for large sample sizes (*i.e.*, large numbers of trials) the binomial probability distribution may be approximated quite accurately by the normal probability distribution,[36] a

---

[33] *Id.*

[34] Such a probability distribution is often displayed in the form of a graph with the possible outcomes (expressed in terms of numbers of successes) plotted along the x-axis and the corresponding probabilities plotted along the y-axis.

[35] Those probabilities, which would be the same for any binomial experiment consisting of 10 trials with the probability of a success on each trial equal to .6, are as follows:

Number of White Chips	Probability
0	0 +
1	.002
2	.011
3	.042
4	.111
5	.201
6	.251
7	.215
8	.121
9	.040
10	.006

[36] The normal probability distribution can be described graphically by the familiar, bell-shaped "normal curve." This curve reaches its highest point at the mean or expected value of the distribution and declines symetrically on each side of the expected value. The curve is concave downward in the vicinity of the expected value, gradually becoming concave upward further away from the expected value (hence the bell-shape). In terms of probabilities, these characteristics of the normal curve mean that: (i) results in the vicinity of the expected value are more probable than those further out on the curve; (ii) results equidistant from the expected value on either side have the same probability of occurrence; (iii) past the point on each side of the expected value where the curve becomes concave upward, the probabilities decline "faster" than the distance from the expected value increases (*e.g.*, a result twice as far from the expected value is less than one-half as likely to occur). The "normal" distribution should not be thought of as the common or usual distribution for observed frequencies, as its name might imply. The prin-

distribution whose properties are well known.[37]

In particular, we can convert the results of a binomial experiment to a variable that has a standard normal distribution,[38] called a Z-statistic or Z-score. The value in this conversion is that the probability associated with any particular Z-score may be obtained from standard tables published in many statistics texts. For the binomial distribution, the Z-score is defined as the difference between the expected and observed outcomes of the binomial experiment divided by its standard deviation. Accordingly, the Supreme Court's suggestion in *Castaneda* that differences between the expected and observed outcomes in excess of two or three standard deviations are statistically significant is equivalent to saying that Z-scores in excess of two or three are statistically significant.[39]

---

cipal significance of the normal distribution in statistics and probability theory is explained *infra* at note 37.

[37] The normal distribution is probably the most important distribution in statistics and probability theory. One reason for this is that (as demonstrated in a set of theorems referred to as "central limit" theorems) under very general conditions sums of non-normal random variables are approximately normally distributed. This makes it possible to make accurate probability calculations for such sums when the exact probability distributions are either unknown or, as in the case of the binomial, difficult to compute. (The number of successes in a binomial experiment can be thought of as the sum of a random variable that for each trial takes on the value 1 for a success and zero for a failure.) *See* MOSTELLER, ROURKE & THOMAS, *supra* note 4, at 275-90, 351-56.

[38] The "standard" normal distribution is the normal distribution whose mean (or expected value) equals zero and whose standard deviation equals one.

[39] The results of a binomial experiment may be considered statistically significant if, under the assumptions of the binomial model, the probability of observing an outcome that differs from the expected value by as much or more than the actual outcome is less than a specified value. This specified value is often set at 1%, 5%, or 10%, although 5% is the most commonly used. *See supra* note 4; *see also* Harper, *Statistics as Evidence of Age Discrimination,* 32 HASTINGS L.J. 1347, 1354 (1981).

The *Castaneda* Court's reference to disparities of more than two to three standard deviations is simply a shorthand way of applying a test of statistical significance. While the specification and interpretation of levels of statistical significance in discrimination litigation pose difficult legal and statistical questions that are beyond the scope of this article, the Supreme Court's implicit discussion of statistical significance in *Castaneda* has engendered some confusion on the subject that merits brief comment. Some courts have interpreted *Castaneda* as establishing two to three standard deviations as the absolute minimum level at which statistical significance may be found and statistical evidence considered probative. *See, e.g.,* EEOC v. Fed. Reserve Bank of Richmond, 698 F.2d 633, 647 (4th Cir. 1983); EEOC v. United Virginia Bank, 615 F.2d 147, 152 (4th Cir. 1980); Cormier v. P.P.G. Industries, Inc., 519 F. Supp. 211, 27 Empl. Prac. Dec. (CCH) ¶ 32,204, 22,608 (W.D. La. 1981); Reynolds v. Sheet Metal Works, 22 Empl. Prac. Dec. (CCH) ¶ 30,738, 14,816 (D.D.C. 1980); Gay v. Local No. 30, 489 F. Supp. 282, 311 (N.D. Cal. 1980), *aff'd,* 694 F.2d 531 (9th Cir. 1982). This is incorrect for a number of reasons. First, as a careful reading of the Court's footnote in *Castaneda* demonstrates, the two to three standard deviations standard was offered as a sufficient, not a necessary, condition for statistical significance. *See* Castaneda v. Partida, 430 U.S. at 496-97 n.17. The disparities there were of such magnitude that the Court was not required to reach, and did not decide, the question of what number of standard deviations would be necessary as a minimum to support a finding of statistical significance.

Second, as a matter of statistical interpretation, statistical significance is not an all-or-nothing proposition; and, therefore, the specification of any particular cut-off point for

The binomial test as applied in *Castaneda* may now be summarized in terms of a set of equations using certain standard symbols to represent the variables discussed above. If we let,

n = the number of binomial trials in a particular experiment,
p = the probability of a success on each trial,
E = the expected number of successes,
O = the observed number of successes,
SD = the standard deviation,

then the statistical analysis in *Castaneda* would proceed as follows:

(1) *Determine "n" and "p" for that particular binomial experiment.* In *Castaneda*, "n" was 870, the total number of persons selected for grand jury duty, and

---

significance, whether expressed in terms of numbers of standard deviations or probabilities, is essentially arbitrary. Accordingly, results that fall just short of whatever level is being used to establish "significance" should not be rejected out of hand and certainly not, as some courts have done, be treated as positive evidence of no discrimination. *See, e.g.,* United States v. Virginia, 454 F. Supp. 1077, 1090 (E.D. Va. 1978), *aff'd in part, rev'd in part,* 620 F.2d 1018 (4th Cir. 1980), *cert. denied,* 449 U.S. 1021 (1980); Garrett v. R.J. Reynolds Indus., 81 F.R.D. 25, 19 Fair Empl. Prac. Cas. (BNA) 13, 19, 23-34 (M.D.N.C. 1978). For example, even if the 5% level is used in the first instance as the threshold for statistical significance, a result that is significant at the 6%, 7%, or 10% level should be regarded as suggestive, entitled to some (although not dispositive) weight, and considered in light of the other evidence — both statistical and nonstatistical — in the case. Some courts have taken this approach. *See* Vuyanich v. Republic Nat'l Bank, 505 F. Supp. 224, 24 Fair Empl. Prac. Cas. (BNA) 128, 165, 223-34 (N.D. Tex. 1980) and cases cited. The problem of imposing arbitrary cut-off points for statistical significance is exacerbated by the apparent practice of some courts of using the upper end of the Supreme Court's suggested range, three standard deviations, as the minimum requirement for statistical significance. *See, e.g.,* Movement for Opportunity and Equality v. General Motors, 622 F.2d 1235, 1259 (7th Cir. 1980); Younger v. Glamorgan Pipe and Foundry Co., 21 Empl Prac. Dec. (CCH) ¶ 30,406, 13,308 (W.D. Va. 1979); Garrett v. R.J. Reynolds Indust. 81 F.R.D. 25, 19 Fair Empl. Prac. Cas. (BNA) 13, 19 (M.D.N.C. 1978). *See also* Department of Labor, Office of Federal Contract Compliance Programs, Order No. 760a1 (March 10, 1983) (suggesting that for disparate treatment analysis statistical disparities must be " 'gross' (in excess of five or six standard deviations). . . . "). Three standard deviations correspond to a level of statistical significance of less than 0.3%, which is below even the most stringent of the standard levels of 1%. Rejecting as statistically "insignificant" any difference between the observed and expected results of a selection process that does not exceed three standard deviations is supported neither by standard social science practice nor by the Supreme Court's language in *Castaneda.*

Third, looking only at the number of standard deviations tends to make statistical significance the only criterion for statistical evidence and ignores the magnitude of the observed disparity. A large disparity may not be "statistically significant" under a statistical model and test because there are too few cases to establish statistical significance. With such small samples, however, the fact that a disparity is not statistically significant does not necessarily mean that it is not "practically significant" in the sense that it may be of a size that is considered important and indicative of possible discrimination. Conversely, a small disparity that is of little or no practical significance may nevertheless be shown to be "statistically significant" if it is based upon a sufficiently large sample of cases. For example, with very large samples a disparity in selection rates between two groups of less than 1% could be found to be statistically significant. It is questionable, however, whether for most purposes such a small disparity would be deemed to be of any practical importance. *See generally* MOSTELLER, ROURKE & THOMAS, *supra* note 4, at 307; BALDUS & COLE, *supra* note 4, at § 9.41.

"p" was .791, reflecting the fact that Mexican-Americans made up 79.1% of the population from which grand jurors were drawn.[40]

(2) *Calculate the expected number of successes* — by multiplying "n" times "p":

$$E = n \times p$$

In *Castaneda*, the expected number of Mexican-American jurors was 688:[41]

$$E = 870 \times .791 = 688$$

(3) *Calculate the standard deviation* — for the binomial distribution associated with that "n" and "p" by taking the square root of the product of the number of trials, the probability of a success, and the probability of a failure:

$$SD = \sqrt{n \times p \times (1-p)}$$

In *Castaneda*, the standard deviation was approximately 12:[42]

$$SD = \sqrt{870 \times .791 \times .209} = \sqrt{143.83} = 11.99$$

(4) *Calculate the Z-score* — that is, the number of standard deviations by which the number of successes actually observed differs from the number of expected:

---

[40] 430 U.S. 482, 496 n.17. The determinations of "n" and "p" are obviously of critical importance to the binomial analysis. While such determinations are fairly straightforward in the coin-tossing and chip-drawing models discussed in the text, in discrimination litigation they can be matters of some complexity and are often the subject of considerable dispute between the parties. In *Hazelwood,* the Supreme Court dealt with two common problems in this area. First, with respect to defining the relevant number of binomial trials, *i.e.* "n," the Court made clear that only hiring decisions made by the Hazelwood School District after it became subject to Title VII in 1972 could be considered in the analysis. 433 U.S. at 309-10. The court of appeals had simply considered the racial composition of Hazelwood's teacher population, which included persons hired prior to, as well as after, 1972. *Id.* at 308. The significance of this for the binomial model is that "n" should be defined in terms of a "flow" figure, that is as the number of persons selected in some sense — *e.g.,* hired, fired, picked for a jury, *etc.* — during a legally relevant period. Defining "n" in terms of a "stock" figure, that is, an aggregate population figure like the teacher population in *Hazelwood,* without reference to the selections made during some specified period will generally be incorrect. Second, the Court in *Hazelwood* made clear that the appropriate way of defining "p," the relevant probability of success on each binomial trial, is to determine the percentage of blacks in a properly defined, qualified labor pool. *Id.* at 308 n.13. The demographic characteristics of this pool, and hence the "p" of the binomial model, will vary, often considerably, depending on both the qualifications that are deemed necessary for the job in question and the geographic areas from which the employer is deemed to draw potential employees. Both of these factors were in dispute in the lower courts in *Hazelwood* and are discussed by the Supreme Court in its opinion. In addition, the Court indicated that the percentage of blacks in the actual applicant pool would also provide a relevant benchmark for comparison with an employer's hiring of blacks and hence an alternative measure of the appropriate "p" for the binomial analysis. *Id. See also* New York City Transit Authority v. Beazer, 440 U.S. 568, 586 (1979). *But see* Dothard v. Rawlinson, 433 U.S. 321, 330 (1977) (applicant data not required to test discriminatory effect on women of height and weight requirements for prison guard positions since requirements themselves may have discouraged otherwise qualified women from applying for those positions).

[41] 430 U.S. 482, 496 n.17.

[42] *Id.*

$$Z = \frac{E - O}{SD}$$[43]

In *Castaneda*, the observed number of Mexican-Americans was 339,[44] so the Z-score was:

$$Z = \frac{688 - 339}{11.99} = 29.1$$

(5) *If the Z-score is greater than 2 or 3, regard the conclusion that the disparity observed was the result of random fluctuations as suspect.* In *Castaneda*, the Z-score was so large that under the assumptions of the binomial model the likelihood of observing such a small number of Mexican-Americans (if selections were made randomly with respect to race) was practically nil.[45] The Court therefore concluded that the statistical disparities established a *prima facie* case of intentional discrimination against Mexican-Americans in that selection process.[46]

III. A Case of Unexamined Assumptions: The Use and Misuse of the Binomial Model in Discrimination Litigation

The Supreme Court's reliance on formal statistical techniques in *Castaneda* and *Hazelwood* to assess the probative value of data evidencing possible discrimination has naturally been emulated by lower courts and litigants.[47] The resulting increase in the accuracy of the analysis of statistical evidence in many discrimination cases has been one of the positive, and was clearly one of the intended, consequences of the Court's use of those techniques in the *Castaneda* and *Hazelwood* opinions.

Furthermore, the particular test employed by the Supreme Court — the binomial — has a number of advantages for use in a litigation context. First, the binomial model is relatively easy to understand and apply. The examples of coin-tossing and drawing chips from a bowl provide homely, but realistic

---

[43] In the actual definition of the Z-score for the binomial distribution, the numerator is defined as the observed number minus the expected number. *See, e.g.*, MOSTELLER, ROURKE & THOMAS at 288. We have reversed the order of the terms in the numerator here and elsewhere in this article. This reversal, which has the effect of changing the sign — but not the magnitude — of the Z-scores, was done to produce positive Z-scores when the observed number of selections of the particular group being examined is less than the expected number (which will be the situation in most of the examples and cases we consider). It is convenient for purposes of discussing the binomial and other statistical models in the context of the language of the Supreme Court to refer to positive rather than negative Z-scores. This change in the sign of the Z-score has no effect whatsoever on the statistical analysis. Because the standard normal distribution, to which the Z-score refers, is symmetric about zero, the statistical significance of any Z-score depends solely on its absolute value and not on its sign (*i.e.*, a Z-score and its negative correspond to exactly the same level of statistical significance).

[44] 430 U.S. 482, 496 n.17.

[45] *Id.* at 496.

[46] *Id.*

[47] *See infra* note 50.

models of binomial experiments.[48] And the often arcane concepts of probability, expected values, and even statistical variability and standard deviations are relatively understandable in the context of the binomial model. Second, the calculations required are fairly simple. Only three variables are involved (the number of trials, the probability of a "success" on each trial, and the observed number of successes), and the most complicated computation is taking a square root — a function now available on inexpensive pocket calculators. Litigants, though well-advised to have expert assistance, can themselves perform the calculations under the varying assumptions about the size of the eligible pool, the number of legally relevant selections, and any other relevant characteristics of the selection process. Third, the binomial model may be applied to a large number of "selection" situations, such as the selection of jurors or the hiring of employees, that are often the subject of discrimination litigation.[49] Accordingly, the binomial model has become the statistical test of preference in discrimination litigation and has been used by courts in a wide variety of cases.[50]

---

[48] *See supra* notes 28-33 and accompanying text.

[49] One limitation of the binomial model, however, is that it may only be applied to selection situations that produce dichotomous results, *e.g.*, called for grand jury duty/not called; hired/not hired; promoted/not promoted. In cases involving continuous or interval variables, such as employee salaries or changes in salaries, or scores on tests or rating systems, statistical techniques other than the binomial must be used to test whether differences between groups with respect to such variables support an inference of discrimination. *See* Baldus & Cole, *supra* note 4, at 12-13. One particular statistical technique that has been used in recent years in an increasing number of such cases is multiple regression analysis. *See, e.g.*, Trout v. Hidalgo, 517 F. Supp. 873, 884-87 (D.D.C. 1981); Segar v. Civiletti, 508 F. Supp. 690, 696-99 (D.D.C. 1981); Vuyanich v. Republic Nat'l Bank, 505 F. Supp. 224, (N.D. Tex. 1980); *see generally* Finkelstein, *The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases,* 80 COLUM. L. REV. 737 (1980). Multiple regression analysis permits an estimate of the average difference between groups on a continuous variable, like salary, after accounting for differences between members of the groups in certain characteristics that are likely to affect that variable, such as (with respect to salary) years of experience and years of education. Multiple regression analysis is more complex both conceptually and computationally than a binomial analysis. But if properly applied and interpreted with the assistance of expert testimony, multiple regression analysis can be a useful tool in discrimination litigation for determining whether observed differences between groups can be attributed, in whole or in part, to the nondicriminatory factors that are included in the analysis. *See* Fisher, *Multiple Regression in Legal Proceedings,* 80 COLUM. L. REV. 702 (1980); Note, *Beyond the Prima Facie Case in Employment Discrimination Law: Statistical Proof and Rebuttal,* 89 HARV. L. REV. 387 (1975).

[50] Williams v. New Orleans Steamship Ass'n, 673 F.2d 742 (5th Cir. 1982), *reh'g denied,* 668 F.2d 412 (5th Cir. 1982) (job assignments for longshoremen); Rivera v. Wichita Falls, 665 F.2d 531 (5th Cir. 1981) (testing of police recruits); Chisholm v. U.S. Postal Service, 665 F.2d 482 (4th Cir. 1981) (promotions of postal workers); Wilkins v. University of Houston, 654 F.2d 388 (5th Cir. 1981), *reh'g denied,* 662 F.2d 1156 (5th Cir. 1981), *cert. denied,* 103 S. Ct. 51 (1982) (faculty wage levels); EEOC v. American Nat'l Bank, 652 F.2d 1176 (4th Cir. 1981), *reh'g denied,* 680 F.2d 965 (4th Cir. 1981), *cert. denied,* 103 S. Ct. 235 (1982) (hiring: binomial applied to applicant flow data); Hameed v. Iron Workers, 637 F.2d 506 (8th Cir. 1980) (admission to apprentice programs); Board of Educ. v. Califano, 584 F.2d 576 (2d Cir. 1978), *aff'd sub nom.* Board of Educ. v. Harris, 444 U.S. 130 (1979) (school assignments of teachers); Otero v. Mesa County Valley School Dist., 568 F.2d 1312 (10th Cir. 1977) (hiring of teachers and support per-

The reliability of the binomial model, however, depends on certain assumptions being met about the selection process being studied. The Supreme Court in *Castaneda* and *Hazelwood* did not make clear what these assumptions are, or how they limit the applicability of the model. A problem arises because these assumptions — of fixed probabilities, only two outcomes, and independence[51] — are satisfied in some but not all of the selection situations in which the binomial model would appear to apply, *i.e.*, in which an "n" and a "p" can be identified and the binomial calculations outlined above can be performed. As a result, the binomial can be and has been used by courts and litigants to test for discrimination in cases where it is inappropriate and produces incorrect results. In the following sections we discuss the assumptions underlying the binomial model, explore the problems that arise in discrimination litigation when they are not met, and suggest alternatives to or modifications of the binomial in such cases.

A. *The Probability of Selecting a Member of the Group in Question is Fixed*

1. Cases in Which This Condition May be Violated

The first condition assumed to be met for the appropriate application of the binomial model requires that the particular class assertedly underrepresented constitute a fixed percentage of the pool of eligible persons from which selections are made. In terms of the formulae described above, this means that the probability "p" that a class member will be chosen each time a selection is made can be specified at the outset of the selection process and does not change throughout the process.

This condition is satisfied for all practical purposes in many types of selection processes.[52] For example, in most jury selection and hiring cases (as in

---

sonnel); Inmates of the Nebraska Penal and Correctional Complex v. Greenholtz, 567 F.2d 1381 (8th Cir. 1977) (granting of parole to prisoners); Vuyanich v. Republic Nat'l Bank, 505 F. Supp. 224 (N.D. Tex. 1980) (hiring of professional, managerial, and clerical workers); Bryan v. Koch, 492 F. Supp. 212 (S.D.N.Y. 1980), *aff'd,* 627 F.2d 612 (2d Cir. 1980) (closing of municipal hospitals); Taylor v. Teletype Corp., 475 F. Supp. 958 (E.D. Ark. 1979), *modified,* 648 F.2d 1129 (8th Cir. 1981), *cert. denied,* 454 U.S. 969 (1981) (demotions and layoffs); Younger v. Glamorgan Pipe & Foundry Co., 21 Empl. Prac. Dec. (CCH) ¶ 30,406 (W.D. Va. 1979) (transfers of employees among job departments); Cooper v. Univ. of Texas at Dallas, 482 F. Supp. 187 (N.D. Tex. 1979), *aff'd,* 648 F.2d 1039 (5th Cir. 1981) (faculty hiring); Rich v. Martin Marietta Corp., 467 F. Supp. 587, 20 Empl. Prac. Dec. (CCH) ¶ 30,111 (D. Colo. 1979) (promotions).

[51] *See supra* notes 28, 29 and accompanying text.

[52] This condition is strictly satisfied only when sampling from an infinite population or sampling with replacement. The latter requires that each person who is selected be put back in the pool and be eligible for all future selections. In the example of drawing poker chips from a bowl, sampling with replacement would mean that after a chip is drawn (and its color recorded), it is immediately put back into the bowl before the next selection is made. Obviously in sampling with replacement, the probabilities of success and failure remain the same for each trial, since the composition of the eligible population is constant throughout the experiment. Most selection situations to which the binomial is applied in discrimination litigation do not involve sampling with replacement, *i.e.*, the persons selected (*e.g.*, for a grand jury, for a job, *etc.*), are removed, at least for a certain time, from the eligible pool.

Castaneda and *Hazelwood*), the number of persons selected is in most cases only a very small fraction of the total pool of persons eligible for selection. In these circumstances, the percentage of class members — for example, blacks — in the eligible pool may be treated as effectively constant, despite the fact that minute changes in the racial composition of the pool occur as persons selected are withdrawn. Nevertheless, for other selection processes such as promotions, layoffs, or testing, the number of persons "selected" (*i.e.*, promoted, laid off, or passing the test) is frequently a not insignificant fraction of the total eligible pool. In these circumstances, the percentage of blacks in the eligible pool changes as selections are made, and the assumption of the binomial model of a constant "p" is not met. In such a case, it is not appropriate to use the binomial model to test for the statistical significance of an observed underrepresentation of class members among the persons selected. In particular, if blacks, for example, are selected at a disproportionately low rate, the percentage of blacks remaining in the eligible pool will tend to increase as selections are made. This in turn will cause the binomial model to err on the "conservative" side from a plaintiff's point of view, *i.e.*, a statistical analysis based on that model will overstate the likelihood that differences in selection rates could be attributed to chance and understate the statistical significance of the racial disparities observed.

This conservative bias inherent in the binomial model under the circumstances described above can be demonstrated in a simple hypothetical. Assume we are analyzing two cases each involving the selection of 100 persons from a population that is 50 percent white and 50 percent black. In the first case the total population from which the selections are to be made is 200,000 persons, but in the second case the total population is only 200 persons.

In the first case, the number of persons selected, 100, is so small compared to the total population of 200,000 that the racial composition of the population will be essentially unaffected by the selection process itself. Thus, the probability of selecting a black on the last "drawing" will be approximately the same as it was on the first "drawing," namely .5. But in the second case, in which half of the persons eligible are selected, the racial composition of the population from which the selections are being made will change over the course of the selection process as those persons selected are withdrawn from the eligible pool. In this case it is likely that the probability of selecting a black in the last drawing will be different — and depending on the results of the selection process, possibly quite different — from .5.

For example, consider the most extreme situation in which there have been 99 persons selected, all of whom are white. What is the probability that the 100th person selected will be black? The answer to this question obviously depends on whether those selections were drawn from the large or the small eligible pool. In the former case, after those 99 selections have been made, there are 199,901 persons left in the pool of whom 99,901 are white and 100,000 are black. Thus, the probability of selecting a black on the next draw-

ing has shifted only slightly from .5 to .5002 (= 100,000/199,901). And, of course, for all the preceding selections the probability of selecting a black was even closer to the original probability of .5. Thus, the deviation from the assumption of a constant "p" for each selection is negligible in this case, and the binomial model will provide accurate estimates of the statistical significance of the observed underrepresentation of blacks among the selectees.

On the other hand, in the case of the smaller population, after those 99 white selections there are 101 persons left in the pool, of whom one is white and 100 are black. Thus, the probability of selecting a black on the 100th draw has increased from .5 to about .99 (= 100/101). And, of course, the probabilities of selecting a black on preceding draws, while smaller than .99, were in many cases also much larger than .5.[53] Obviously, the assumption underlying the binomial model that each time a selection is made the chance of selecting a black is the same has been violated, and as a result, the binomial model will not provide a reliable means of assessing the likelihood that an observed underrepresentation of blacks is simply the result of random fluctuations inherent in the sampling process.

As indicated above, if a binomial analysis is applied to the selection data for the smaller population and the initial black representation in the eligibility pool of 50 percent is used as the fixed probability of selecting a black throughout the selection process (*i.e.*, if .5 is used as the "p" in the binomial formulae), the underrepresentation of blacks will appear to be less statistically significant than it actually is. In the extreme case we have been considering, in which 100 whites and zero blacks were selected, this bias will make little practical difference because the racial disparities will in any event appear very highly significant. In closer cases, however, such inappropriate use of the binomial model could make statistically significant results appear insignificant or only marginally significant.[54]

2. Alternatives to the Binomial Model — Tests of Differences Between Proportions

Since the binomial assumption of a fixed probability of selection on each trial fails when the number of persons selected is a substantial fraction of the eligible population, an alternative statistical test must be used in those circumstances. Such alternative tests involve looking at the selection data in a slightly different fashion from that used in the binomial.

In the selection situation examined previously, we looked at the division of selectees into racial or ethnic groups and applied the binomial model to com-

---

[53] For example, assuming that all previous selections had been white, the probability of selecting a black on the 99th draw would be about .98, on the 50th draw about .66, and on the 25th draw about .57.

[54] *See infra* notes 85-87 and accompanying text.

pare the number of those selected of a given group with an expected number of selectees from that group. The logic of this approach is that if, for example, blacks represent X percent of the eligible pool we would expect in a racially neutral system that they would represent approximately X percent of those selected. The same data can also be looked at by considering the selection percentages for each group from its own eligible population. The logic of this approach is that if, for example, Y percent of the eligible whites are selected, we would expect in a racially neutral system that approximately Y percent of the eligible blacks would also be selected. These selection percentages can be compared, not by use of a binomial test, but by use of tests designed to test differences between proportions.

One such test is referred to as the "normal theory test" of the difference between two proportions. Like the Z-score test for the binomial model,[55] this test relies on a normal approximation to a distribution—in this case the distribution of the difference between the two proportions.[56] The difference-in-proportions test is usually expressed in terms of the following standard notation:

p_0 = proportion of the total eligible population selected
p_1 = proportion of the first group selected
p_2 = proportion of the second group selected
n_1 = the number of persons of the first group in the eligible population
n_2 = the number of persons of the second group in the eligible population

The difference in proportions, *i.e.*, selection rates, for the two groups, $p_1 - p_2$, is tested by calculating a Z-statistic as follows:

$$Z = \frac{p_2 - p_1}{\sqrt{p_0 (1-p_0) (1/n_1 + 1/n_2)}}$$

The Z-statistic here has approximately a standard normal distribution, so that the same procedure is used in conducting the test as in the binomial test of *Castaneda*.[57]

For example, consider a situation in which 100 persons 12 black and 88 white, have been selected from an eligible population of 200 persons, 40 black and 160 white. These data are set out below.

---

[55] *See supra* text accompanying notes 38-39.

[56] This test also requires that there only be two groups and two possible outcomes for each group (*e.g.*, blacks and whites/hired and not hired). Furthermore, the selections must be independent of one another, that is, the selection or non-selection of one person cannot influence the outcome of the selection process for any other person. For a discussion of this test, see SNEDECOR & COCHRAN, *supra* note 4, at 124.

[57] In other words, if the Z-score is of a sufficient magnitude, the conclusion that the disparity between the two proportions is the result of random fluctutions is regarded as suspect. *See supra* text accompanying notes 27, 45-46.

Table 1

Hypothetical Selection Process

	Black	White	Total
Selected	12	88	100
Total Eligible	40	160	200

With 50% of the eligible population selected, the binomial cannot be used to determine the probability that in a racially neutral selection system as few as 12 blacks would be chosen out of 100 selections. The difference-in-proportions test, however, can be used to compare the observed proportion of selections from the black eligibles, 12/40 or .30, with the corresponding proportion for whites, 88/160 or .55. In terms of the notation defined above:

$p_1 = 12/40 = .30$
$p_2 = 88/160 = .55$
$p_0 = 100/200 = .50$
$n_1 = 40$
$n_2 = 160$

$$Z = \frac{.55 - .30}{\sqrt{(.5)\,(.5)\,(1/40 + 1/160)}} = \frac{.25}{.088}$$

$Z = 2.83$

This result is significant at the one-percent level and, since the Z-score exceeds 2, under the *Castaneda* standard as well.[58]

Other tests besides the normal theory test of a difference between two proportions may be used in appropriate circumstances. When the sample sizes are small the "hypergeometric" or equivalently "Fisher's Exact" test is used because it is based on exact probability calculations and thereby avoids the normal approximations used in the normal theory test, which are inaccurate for small samples.[59] The "chi-square" test, discussed briefly below, is an equivalent test for a difference between two proportions. (The chi-square test may also be used for differences among more than two proportions.) The choices between these tests are often based on narrow technical grounds and are best made with expert statistical assistance.

3. Misuse of the Binomial in Judicial Decisions

An example of a case in which the improper use of the binomial model may have had decisional significance is provided by Inmates of the *Nebraska Penal and Correctional Complex v. Greenholtz*.[60] In *Nebraska Inmates*, the United States Court of Appeals for the Eighth Circuit used the binomial model ap-

---

[58] *See supra* text accompanying note 27.

[59] *See* note 86 *infra*. For an example of the use of the hypergeometric in a case involving the selection of hospitals for closure, see the discussion of Bryan v. Koch, 492 F. Supp. 212 (S.D.N.Y. 1980), *aff'd*, 627 F.2d 612 (2d Cir. 1980), *infra* at Section IIIC.

[60] 567 F.2d 1368 (8th Cir. 1977), *cert. denied*, 439 U.S. 841 (1978).

proach, as set out in *Castaneda* and *Hazelwood*, to analyze the selection of Nebraska prisoners for parole.[61] A class of Native-American and Mexican-American inmates of the Nebraska Penal and Correctional Complex had brought an action under section 1983[62] alleging that the Nebraska Board of Parole had denied them discretionary parole on racial and ethnic grounds in violation of the fourteenth amendment to the Constitution of the United States.[63] A major part of the plaintiffs' proof involved a statistical showing that the Native-American and Mexican-American inmates who were eligible for discretionary parole received such parole at substantially lower rates than did eligible white and black inmates.[64] The data on granting of discretionary parole, stratified by racial and ethnic groups, were as follows:[65]

Table 2

	Inmates Eligible For Parole		Inmates Paroled[66]		
	Number	Percentage of Total Eligibles	Number	Percentage Paroled	Percentage of Total Parolees
	(1)	(2)	(3)	(4)	(5)
White	590	65.4	358	60.7	66.9
Black	235	26.1	148	63.0	27.7
Native American	59	6.5	24	40.7	4.5
Mexican-American	18	2.0	5	27.8	0.9
TOTAL	902	100.0	535	59.3	100.0

[61] *Id.* at 1375-79.

[62] 42 U.S.C. § 1983.

[63] 567 F.2d 1368 (8th Cir. 1977), *cert. denied*, 439 U.S. 841 (1978).

[64] *Id.* at 1371-73. Since this was a § 1983 claim alleging discrimination in violation of the Constitution, the plaintiffs had to prove that they were the victims of intentional discrimination; mere disparate impact without discriminatory purpose would be insufficient to make out a constitutional violation. Washington v. Davis, 426 U.S. 229, 239-48 (1976). "Disparate impact is not irrelevant, but is not the sole touchstone of an invidious racial discrimination forbidden by the Constitution." *Id.* at 242. In contrast with the constitutional prohibitions on discrimination, Title VII of the Civil Rights Act of 1964 prohibits both intentional discrimination in employment and employment practices that have a disparate impact on a group protected by the statute and that are not justified by business necessity even if adopted without discriminatory intent. *See supra* note 12. However, statistical proof that the challenged governmental action bears more heavily on some races or ethnic groups than others will generally be an important element in the proof of purposeful discrimination, at least in class actions. *Id.* at 239-48. Moreover, when the racial disparities shown by such proof are severe enough, they alone may support an inference of intentional discrimination, *Id.*; Hazelwood School Dist. v. United States, 433 U.S. at 307-08. Accordingly, in most cases involving classwide claims of purposeful discrimination, plaintiffs include statistical proof of disparate impact. Conversely, if the defendants can show that the challenged governmental action does not have a significant disparate impact on a particular class, they will in most instances defeat any claim that such action was undertaken with the intent to discriminate against that class.

[65] *See Nebraska Inmates*, 567 F.2d at 1371.

[66] The percentages in column 4 are derived by dividing the numbers in column 3 by the

The district court's analysis in *Nebraska Inmates*[67] of these data is worth reviewing briefly because it is typical of the statistically unsophisticated approach common in judicial decisions in discrimination cases prior to the Supreme Court's endorsement of formal statistical hypothesis testing in *Castaneda* and *Hazelwood*.[68] The district court's analysis consisted of a simple arithmetic comparison of the percentages listed above in column (2) and column (5): that is, the court compared the percentage that each group represented in the eligible population with that group's percentage representation in the population of parolees.[69] For example, the court noted that while 65.4% of the 902 inmates eligible for parole were white, 66.9% of the 535 inmates receiving parole were white, a difference of only 1.5%[70] The corresponding percentage disparities for blacks, Native-Americans, and Mexican-Americans were in each case 2% or less, which the court found to be an insignificant difference.[71] Accordingly, the court concluded that the plaintiffs' statistical evidence was not probative of disparate impact and contributed nothing toward making out a case of purposeful racial and ethnic discrimination in Nebraska's discretionary parole process.[72]

The district court's analysis in *Nebraska Inmates* of the statistical evidence is defective in at least two important respects. First, the simple comparison of percentages undertaken by the Court is insensitive to the size of the populations to which the percentages refer. Statistical theory demonstrates that the larger the sample size, the more statistically significant any particular absolute difference in such percentages is, all other things being equal.[73]

Second, simply comparing the percentages a minority group represents in the pre- and post-selection populations as the district court did, can be misleading since as the former percentage diminishes, increasingly smaller differences between the two percentages can become significant.[74] In fact, using the

---

corresponding numbers in column 1. The percentages in column 5 are derived by dividing each of the numbers in column 3 by the number 535, the total of column 3.

[67] 436 F. Supp. 432 (D. Neb. 1976), *rev'd*, 567 F.2d 1368 (8th Cir. 1977), *cert. denied*, 439 U.S. 841 (1978).

[68] The district court's decision, which was handed down in 1976, predated *Castaneda* and *Hazelwood*, which were decided during the summer of 1977.

[69] *Id.* at 439-42.

[70] *Id.* at 441-42.

[71] *Id.*

[72] 432 F. Supp. at 441-42. And, after analyzing the plaintiffs' nonstatistical evidence, the court held that the plaintiffs had failed to establish a *prima facie* equal protection case and entered judgment for the defendants. *Id.* at 442.

[73] For example, the percentages the *Nebraska Inmates* district court compared would be the same whether the numbers of eligible and paroled inmates in each racial and ethnic group were only one-tenth as large or ten times as large as those shown in Table 1. Yet the racial and ethnic underrepresentation revealed in those percentage disparities may be insignificant when a total of only 90 inmates are involved, but quite significant when the population numbers 9,000. But there is no systematic, statistically reliable way to take variations in sample size into account when a simple arithmetic comparison of percentages is used as the measure of statistical disparity.

[74] For example, consider a situation in which two minority groups are subject to a selec-

district court's approach, it is impossible to show disparate impact for any minority group whose percentage of the eligible population is less than whatever minimum difference in percentages is set as a threshold for establishing significance. For example, as the court of appeals in *Nebraska Inmates* pointed out, since Mexican-Americans represented only 2.0 percent of the eligible population, even if zero Mexican-Americans had been paroled, the statistical shortfall would have been only 2.0 percent, which the district court held to be insignificant.[75] Therefore, even if the Nebraska Parole Board followed a policy of completely excluding Mexican-American inmates from discretionary parole, the district court's test would indicate that the resulting disparity was "insignificant." Thus, the district court's approach in *Nebraska Inmates* has the perverse effect that, all other things being equal, the more "minority" a group is (*i.e.*, the smaller its representation in the relevant eligible population) the more difficult it is to show that a selection process has an adverse impact on that group.[76]

After properly rejecting the district court's treatment of the statistical evidence, the Eighth Circuit in *Nebraska Inmates* undertook its own analysis of the parole data, and in doing so quite naturally attempted to apply the new learning of the *Castaneda* and *Hazelwood* cases. The court reasoned that the 6.5 percent representation of Native-Americans and the 2.0 percent representation of Mexican-Americans in the prisoner population from which persons were selected for discretionary parole were analogous to the 79.1 percent figure that Mexican-Americans represented in the population from which persons were selected for grand jury duty in *Castaneda*.[77] Accordingly, the court reasoned that if the 535 discretionary paroles had been distributed proportionately among all racial and ethnic groups, Native-Americans would have received about 35 (535 × .065) and Mexican-Americans about 11 (535 × .02) of those paroles.[78] Thus, the differences between those "expected" and the "observed" numbers

---

tion process. Assume the first group represents 40% of the eligible pool and receives 35% of the selections, while the second group represents 6% of the eligible pool and receives 1% of the selections. In each case, the statistical disparity under the approach employed by the district court is 5% (40% minus 35% for the first group, 6% minus 1% for the second group). However, for the first group the selection process results in a "shrinkage" in their proportional representation in the relevant populations of only 12.5% (.05/.40), while for the second group the drop from 6% to 1% represents a shrinkage of over 80% in their population share (.05/.06 = .833). When the data are looked at in this fashion, it seems clear that the second group has been more adversely impacted by the selection process than the first group. Whether the adverse impact of the selection process on either group is statistically significant depends on the sizes of both the eligible and the selected populations. However, for any given set of sample sizes the statistical significance of the 5% disparities described above will be greater for the second group than for the first group.

[75] 567 F.2d at 1377 n.18.

[76] The court of appeals also noted that this problem — which is a function of the minority group's small *comparative* size in the *eligible* population — bears no necessary relationship to the problem of small sample size discussed above, which is concerned with the *absolute* size of the *selected* population. 567 F.2d at 1377 n.18.

[77] *Id.* at 1377-79.

[78] *Id.* at 1378.

of paroles were 11 (35 - 24) for Native-Americans and 6 (11 - 5) for Mexican-Americans.[79] Still tracking closely the *Castaneda* model, the appeals court calculated the standard deviations for both groups; and, since the difference between the expected and observed numbers of discretionary paroles was less than two standard deviations in each case, the court concluded that the statistical evidence did not substantially support the plaintiffs' *prima facie* case that racial and ethnic discrimination infected Nebraska's parole decisions.[80]

What the appeals court in *Nebraska Inmates* failed to do was examine the assumptions underlying the binomial model before applying it to the statistical proof on discretionary paroles. The court's conclusion that the percentage figures it plugged into the binomial equations were analogous to the population percentages the Supreme Court used in *Castaneda* and *Hazelwood* to calculate the statistical significance of the ethnic and racial underrepresentations in those cases appears on the surface to be correct. But, unlike the situations in those cases, in *Nebraska Inmates* the number of persons selected was a substantial percentage of the eligible pool: 535 out of 902 or 59.3% of the inmates eligible for discretionary parole received it.[81] Accordingly, it cannot be assumed that the percentages of Native-Americans and Mexican-Americans in the eligible pool remained constant during the selection process. And, in fact, by the end of the selection process, the Native-American percentage had increased to 9.5% and the Mexican-American percentage to 3.5%,[82] which are substantially higher than 6.5% and 2.0% — their respective representation in the eligible pool before any selections were made. Thus, when the appeals court used the latter figures in the binomial formulae as the respective, fixed probabilities of selecting a Native-American or Mexican-American on each of the 535 "draw-

---

[79] *Id.*

[80] *Id.* at 1374. The calculations performed by the Court may be summarized as follows:

(a) Native Americans

$$p = 59/902 = .065$$

Expected number $= 535 \times .065 = 35$

Observed number $= 24$

Standard deviation $= \sqrt{n(p)(1-p)}$

$$= \sqrt{535(.065)(.935)} = 5.72$$

$$Z = \frac{35-24}{5.72} = 1.92$$

(b) Mexican-Americans

$$p = 18/902 = .02$$

Expected number $= 535 \times .02 = 10.7$

Observed number $= 5$

Standard deviation $=$

$$\sqrt{535(.02)(.98)} = 3.24$$

$$Z = \frac{10.7-5.0}{3.24} = 1.76$$

Id.

[81] *Id.* at 1371.

[82] There were 35 Native Americans and 13 Mexican-Americans among the 367 inmates who did not receive parole. *See supra* Table 2.

ings," it was understating the probability of randomly selecting a member of one of these two groups on later drawings and thus biasing its calculations in the direction of finding nonsignificance.

Application of an appropriate statistical test to the data in Nebraska Inmates demonstrates the practical importance of the Eighth Circuit's misapplication of the binomial model. For example, using the difference-in-proportions test (which compares the success rate for one group with that of the rest of the population[83]) for Native-Americans, the results obtained are quite different from those obtained with the binomial. Of the 59 Native-Americans eligible for discretionary parole, 24 or 40.7% received it.[84] The corresponding figures for all non-Native-Americans are 511 out of 843 or 60.6% and for the entire inmate population 535 out of 902 or 59.3%.[85] In terms of the symbols and formulae for this test set out above, these numbers can be expressed as follows:

$p_1 = 24/59 = .407 =$ (the parole rate for Native-Americans)
$p_2 = 511/843 = .606 =$ (the parole rate for non-Native-Americans)
$p_0 = 535/902 = .593 =$ (the parole rate for all inmates)
$n_1 = 59 =$ (the number of Native-Americans eligible for parole)
$n_2 = 843 =$ (the number of non-Native-Americans eligible for parole)

The Z-statistic may now be calculated as follows:

$$Z = \frac{p_2 - p_1}{\sqrt{(p_0)\,(1-p_0)\,(1/n_1 + 1/n_2)}}$$

$$= \frac{.606 - .407}{\sqrt{(.593)(.407)(1/59 + 1/843)}}$$

$$= 3.01$$

A similar calculation for Mexican-Americans yields a Z-score of 2.75.[86] Thus, the expected number of paroles diverges from the observed number of approximately three standard deviations for both Native-Americans and Mexican

---

[83] *See supra* notes 55-58 and accompanying text.
[84] 567 F.2d at 1371.
[85] *Id.*
[86] The calculation for Mexican-Americans may be summarized as follows:

$p_1 = 5/18 = .278$
$p_2 = 530/884 = .600$
$p_0 = 535/902 = .593$
$n_1 = 18$
$n_2 = 884$

$$Z = \frac{.600 - .278}{\sqrt{(.593)(.407)(1/18 + 1/884)}}$$

$Z = 2.75$

If the size of the eligible population (*i.e.*, total number of inmates eligible for parole) were small (say, below 20) or if the smallest expected number of "successes" (*i.e.*, number of paroles) for any group were very small (say, below 5), then the normal approximation might not be adequate, and use of the hypergeometric would be recommended. *See* SNEDOCOR & COCHRAN, *supra* note 4, at 127. Neither condition obtains here, however.

Americans — not slightly less than two standard deviations for both groups as the appeals court mistakenly calculated. Since the court, following the *Castaneda* decision, indicated it would regard differences of two to three standard deviations as statistically significant, it is possible its decision in *Nebraska Inmates* would have been different if the court had not been misled about the statistical significance of the discrepancies in parole rates through its improper use of the binomial analysis.[87]

4. Guidelines for the Use of the Binomial Model: The Size of the Population Selection Rate

There is no bright-line standard for the population selection rate — that is, the total number of persons selected divided by the total number of persons in the population from which selections are made — that separates those selection situations for which the binomial model is accurate from those for which it is inaccurate. Rather, as the population selection rate increases the binomial gradually becomes unreliable. As a rough rule of thumb, however, if the percentage of the eligible population selected is 10 percent or less, the binomial

---

[87] This mistaken use of the binomial model appears in a number of other court decisions, and even in the testimony of experts in discrimination litigation. For example, in Taylor v. Teletype Corp., 475 F. Supp. 958 (E.D. Ark. 1979), *modified,* 648 F.2d 1129 (8th Cir. 1981), *cert. denied,* 454 U.S. 969 (1981), both the plaintiffs' and the defendants' experts incorrectly applied the binomial model to a situation where the number of persons "selected" (in that case demoted and laid off) was a substantial percentage of the pool of persons eligible for selection. 475 F. Supp. at 961-64. The defendants' expert, applying the binomial analysis to data on demotions, calculated a Z-score of 1.4, which the court found did not support any inference of discrimination. *Id.* at 961-62. But the difference-in-proportions test applied to the same data yields a Z-score of 1.84. This difference may have been important since the court appears to have recognized any Z-score above 1.645, which is the 5% level of significance on a "one-tail" test (*see* BALDUS & COLE, *supra* note 4, at 307-08), as probative of adverse impact and supportive of a *prima facie* case of discrimination. The plaintiffs' expert, working with data on layoffs, similarly misapplied the binomial model, although the Z-score resulting from his analysis of 2.23 was considered high enough by the court to support the plaintiffs' claims of discrimination. *Id.* at 962-63; *see also* EEOC v. Fed. Reserve Bank of Richmond, 698 F.2d 633, 650-54 (4th Cir. 1983) (promotions; binomial applied to data on promotions from certain grades where between 37% and 43% of eligible employees received promotions); Williams v. New Orleans Steamship Ass'n, 673 F.2d 742 (5th Cir. 1982), *reh'g denied,* 688 F.2d 412 (5th Cir. 1982) (job assignments; binomial applied to data on assignment of "deck and wharf" jobs to longshoremen where 40% to 50% of the workforce received such assignments); Harrell v. Northern Electric Co., 672 F.2d 444 (5th Cir. 1982), *modified,* 679 F.2d 31 (5th Cir. 1982), *cert. denied,* 103 S. Ct. 449 (1982) (hiring; binomial applied to applicant flow data where 30% of applicants were hired); Wilkins v. University of Houston, 654 F.2d 388 (5th Cir. 1981), *reh'g denied,* 662 F.2d 1156 (5th Cir. 1981), *cert. denied,* 103 S. Ct. 51 (1982) (wage levels; binomial applied to data on "underpaid" employees where 49% of relevant employee pool were in "underpaid" category); Hameed v. Iron Workers, 637 F.2d 506 (8th Cir. 1980) (admittance to union apprenticeship programs; binomial applied to rating system where 96% of applicants achieved cut-off score and to selections for apprentice positions where selection rate was 67%); Davis v. Dallas, 487 F. Supp. 389 (N.D. Tex. 1980) (hiring; binomial applied to applicant-flow data where 18% of the applicants were hired). Marsh v. Eaton Corp., 25 Fair Empl. Prac. Cas. (BNA) 57, 62 n.13 (N.D. Ohio 1979), *aff'd in part and rev'd in part,* 639 F.2d 328 (6th Cir. 1981) (job placements; binomial applied to data on placement of new hires in lower level positions where 82% of new hires placed in those positions).

model will produce reasonably accurate results. If the overall selection percentage exceeds 10 percent, the difference-in-proportions test should generally be used.[88]

As a practical matter, this means the binomial model will generally be applicable in situations where selections are being made from the general population in a particular area, such as the selection of grand jurors in *Castaneda*, or from an area-wide labor pool, such as the hiring of teachers in *Hazelwood*. In such cases, the number of selections being made will rarely be so large as to affect the demographic or work-force statistics from which the probability of a success on each binomial trial, the "p" of the binomial formulae, is derived. When the eligible population is more narrowly circumscribed, such as in an applicant-flow analysis in a hiring or promotion case, care must be taken before applying the binomial since it is possible the selection rate will be substantial. Furthermore, the binomial will rarely be appropriate in cases challenging employment or other types of tests since it is unusual for the passing rate on such tests to be below 10%.[89]

B. *Each Binomial Trial Can Have Only Two Possible Outcomes*

The second condition assumed to be met in applying the binomial model requires that there be only two possible outcomes for each trial. For example, if the model is to be used to analyze the selection rate for blacks, it is necessary to

---

[88] *See* BALDUS & COLE, *supra* note 4 § 9A.12 (1980 & Supp. 1982) (suggesting 10% rule of thumb, while citing other authorities that suggest other figures from 1% to 20%). If the overall selection rate exceeds 90%, the binomial analysis may be applied to data on the "non-selected" population (to test the difference between the expected and observed numbers of nonselected persons of a particular group) since the overall "non-selection" rate will necessarily be below 10%.

[89] It has been argued that the binomial model is never an appropriate means of determining whether pass-fail data demonstrate that an employment test has a statistically significant adverse impact on a particular group. *See* Shoben, *Differential Pass-Fail Rates in Employment Testing: Statistical Proof Under Title VII*, 91 HARV. L. REV. 793 (1978). Shoben acknowledges that the binomial may be used to compare, for example, the racial composition of a sample population with that of the larger population from which it was selected; but she argues that because a test divides the sample population into four relevant categories, two racial groups, each composed of passers and failers, the binomial is inapplicable. *Id.* at 795-96. Sometimes, however, a conditional test is performed in which the persons who pass the test are the persons "selected" by the test. Hence, the total number of passers defines "n," the number of binomial trials, and the percentage of the group in question among the test-takers defines "p," the probability of success on each trial. An expected number of passers for the particular group can be calculated by multiplying n times p, and a standard deviation can be calculated using the usual binomial formula, $\sqrt{n \times p \times (1 - p)}$. The problem with using the binomial model in such a case is that — as indicated in the text — usually a not insubstantial proportion of the persons taking a test pass it, thereby rendering the binomial calculations inaccurate. However, if a particular test had a low overall pass rate (*i.e.*, below 10%), the binomial model would provide as reliable an estimate of the statistical significance of any racial, ethnic, or sexual disparities among those "selected" as it would in a hiring or any other type of selection case with a comparable overall selection rate. However, since as a practical matter the binomial model will not be applicable in most testing cases, we agree with Shoben that the binomial should generally be avoided and the difference-in-proportions test used in its stead when one is examining differential pass-fail rates on tests.

divide the entire population into two categories — blacks and nonblacks. This is fairly straightforward when the model is used to test the effects of a selection process on one particular group, as in *Castaneda* (Mexican-Americans) and *Hazelwood* (blacks). When a selection process allegedly has an adverse impact on two (or more) distinct groups, however, the question arises how the selection experience of one disfavored group should be treated when the data on another disfavored group are being analyzed. For example, if an employer's hiring system appears to have an adverse impact on both blacks and Hispanics, how should Hispanics be counted when the binomial analysis is applied to the black selections (and vice versa when the Hispanic selection rate is analyzed)?[90] Simply including Hispanics in the nonblack category would have the effect of lowering the nonblack selection rate and reducing the perceived disparity in selection rates between blacks and nonblacks. In other words, the impact of the hiring process will appear less adverse to blacks if they are compared with all "nonblacks," including the disfavored Hispanics, than if they are compared only with the favored whites.

There are two possible adjustments to the binomial model that can be made in these circumstances. The first and simplest is to combine the data on blacks and Hispanics and perform the binomial calculations on the selection rate for this combined "minority" category. Such a combination may be appropriate if there is some evidence that the selection system being examined discriminates against both groups in a similar fashion and if both groups were included in the certified class.

However, if the alleged discrimination is different in nature or if the class as certified includes only one of the two groups, it may not be appropriate to combine the data on the two groups. For example, it would seem inappropriate to base a finding of discrimination against blacks on statistical evidence of adverse impact that would be insignificant but for the inclusion of data on Hispanics who are not part of the plaintiff class. This suggests that in these circumstances the binomial model should be adjusted to analyze the selection process as it affects blacks and whites only (*i.e.*, not treat the second disfavored group, Hispanics, as part of either the black or white group for purposes of the analysis).

Both of these possible adjustments can be illustrated through the use of a simple example. Consider a hypothetical employer who hires 100 persons from a labor pool that is 50% white, 25% black, and 25% Hispanic. The racial distribution of the hirees is 64 whites, 18 blacks, and 18 Hispanics. Applying the binomial model to the entire selection process for either blacks or Hispanics would lead to the following calculations:

---

[90] In terms of the physical model of drawing chips from a bowl, this would correspond to having chips of three different colors, *e.g.*, black, brown, and white, in the bowl. If we assume that both black and brown chips are underrepresented among the chips drawn from the bowl, the question considered in the text is when the selection of black chips is being analyzed, how should the brown chips be counted (and vice versa).

Expected number of hires = $100 \times .25 = 25$
Observed number of hires = 18
Standard deviation = $\sqrt{100\ (.25)\ (.75)} = \sqrt{18.75} = 4.33$

$$Z = \frac{25\text{-}18}{4.33} = \frac{7}{4.33} = 1.62$$

This result is not quite significant at the five percent level and is below the two- to three-standard suggested in *Castaneda* as well. If, however, either of the two adjustments to the model discussed above are used, different results are obtained.

First, if blacks and Hispanics are combined into a "minority" category the eligible pool would then be 50% minority and 50% white, and the hires would be 36 minority and 64 white. The Z-score calculation for this adjusted model would proceed as follows:

Expected number of hires = $100 \times .50 = 50$
Observed number of hires = 36
Standard deviation = $\sqrt{100\ (.5)\ (.5)} = 5$

$$Z = \frac{50\text{-}36}{5} = \frac{14}{5} = 2.8$$

Second, the model can be adjusted to compare only blacks versus whites. This can be done by removing the Hispanics from *both* the labor pool and the group of persons selected and then recalculating the expected value of black hires and the standard deviation. Blacks represent one-third (.25/.75) of this adjusted black-white eligible pool and 22% (18/82) of the black-white selections. Accordingly, the binomial calculations after these adjustments are made would be as follows:[91]

Expected number of hires = $82 \times .33 = 27.1$
Observed number of hires = 18
Standard deviation = $\sqrt{82\ (.33)\ (.67)} = 4.26$

$$Z = \frac{27.1 - 18}{4.26} = 2.14$$

The results with these adjusted models are significant at the 1% and 5% levels, respectively, and — since the difference between expected and observed values in each case is more than two standard deviations, — under the *Castaneda* standard as well. Thus, a selection process that adversely affects both blacks and Hispanics as compared, either jointly or separately, to whites could under a simple, unadjusted binomial analysis appear not to have a statistically significant adverse impact on either one.[92]

---

[91] Since in this example the data for Hispanics mirror exactly those for blacks, the adjusted analysis for Hispanics obviously would produce results identical to the results for blacks described in the text.

[92] In these types of cases, the plaintiffs or the court must identify the nature of the discriminatory process they are reporting on and testing. Different kinds of discrimination will call for different statistical comparisons between and among groups.

The *Nebraska Inmates*[93] decision discussed above provides an example in an actual judicial decision of a failure to consider this feature of the binomial model. The plaintiffs there alleged discrimination against two distinct groups — Native Americans and Mexican-Americans — and the appeals court performed separate binomial analyses to test for discrimination against each one. When it analyzed the parole experience of Native-Americans, however, the court included Mexican-Americans in the comparison group of non-Native-Americans who were supposedly not discriminated against, and *vice versa* when it analyzed the parole rate for Mexican-Americans. As explained above, this has the effect of reducing the apparent disparity in parole rates between each of these disfavored groups and the rest of the inmate population. Thus, even if the binomial had otherwise been applicable to the parole data in *Nebraska Inmates*, the failure of the circuit court to adjust the binomial calculations to take into account the fact that discrimination was being alleged against two groups would lead to an underestimation of the statistical significance of the observed disparities.[94]

Furthermore, the difference-in-proportions test, which as discussed above may be applied in certain selection situations where the binomial is not available (because the probabilities of selection are not the same for each trial), shares with the binomial the assumption that the population being studied consists of only two groups. And, therefore, the same adjustments required for the binomial analysis when discrimination against two or more distinct groups is alleged — that is, either combining the two disfavored groups or eliminating the experience of all other disfavored groups from the selection data when the selection rate for one disfavored group is being analyzed — must be made in order to apply the difference in proportions test in such circumstances as well.

For example, in the *Nebraska Inmates* case if the difference-in-proportions test is applied to a combined Native-American and Mexican-American group, the Z-score for the disparity in parole rates between that group and the rest of the inmate population is 4.08, a highly significant result.[95] Since in *Nebraska In-*

---

[93] Inmates of the Nebraska Penal and Correctional Complex v. Greenholtz, 567 F.2d 1368 (8th Cir. 1977), *cert. denied,* 439 U.S. 841 (1978); *see supra* notes 60-87 and accompanying text for the discussion of *Nebraska Inmates.*

[94] If the binomial calculations had been adjusted so as to compare Native Americans and Mexican-Americans separately with only blacks and whites, the Z-scores would have increased from 1.92 to 1.98 for Native Americans and from 1.76 to 1.81 for Mexican-Americans. If Native Americans and Mexican-Americans had been combined into one group, the binomial calculations would have yielded a Z-score of 2.58 for the shortfall of parolees in that combined group.

[95] The difference-in-proportions calculations for the combined Native American/Mexican-American class would proceed as follows. The data are derived from Table 2, *supra*, and the symbols are those defined *supra* at text accompanying notes 55-56.

$$p_1 = 29/77 = .377$$
$$p_2 = 506/825 = .613$$
$$p_0 = 535/902 = .593$$
$$n_1 = 77$$

mates the Native-Americans and Mexican-Americans were part of the same plaintiff class and the alleged discrimination against both groups arose in part, according to plaintiffs, from a common source — a supposed ignorance and insensitivity on the part of parole board members to the foreign cultures and values of these two groups[96] — combining the parole data for both groups arguably would have been appropriate in the circumstances of that case.

Even if the plaintiffs were charged with proving discrimination against each group separately, however, the difference-in-proportions test should be adjusted so that each disfavored group is tested separately against the rest of the inmate population. For example, to test the underrepresentation of Native-Americans, the Mexican-Americans should first be taken out of the data base and new population figures determined for a revised "non-Native-American" class consisting solely of blacks and whites. This yields new (and higher) selection rates for non-Native-Americans and for the population as a whole. A comparable adjustment could be made when analyzing the Mexican-American data. When this is done, the Z-score for the disparity in selection rates for Native-Americans is 3.12 and that for Mexican-Americans is 2.88, slightly higher, and thus slightly more statistically significant, than the results calculated earlier when all the data, including those for the other disfavored group, were used in each analysis.[97] These results give a better assessment of

---

$$n_2 = 825$$

$$Z = \frac{.613 - .377}{\sqrt{(.593)(.407)(1/77 + 1/825)}}$$

$$Z = 4.03$$

[96] 567 F.2d 1368, 1370.

[97] The difference-in-proportions calculations with the adjustments discussed in the text would proceed as follows. The data are derived from Table 2, *supra,* and the symbols are those defined *supra* at text accompanying notes 56-57.

(a) Native Americans v. Whites and Blacks

$$p_1 = 24/59 = .407$$
$$p_2 = 506/825 = .613$$
$$p_0 = 530/844 = .600$$
$$n_1 = 59$$
$$n_2 = 825$$

$$Z = \frac{.613 - .407}{\sqrt{(.6)(.4)(1/59 + 1/825)}}$$

$$Z = 3.12$$

(b) Mexican-Americans v. Whites and Blacks

$$p_1 = 5/18 = .278$$
$$p_2 = 506/825 = .613$$
$$p_0 = 511/843 = .606$$
$$n_1 = 18$$
$$n_2 = 825$$

$$Z = \frac{.613 - .278}{\sqrt{(.606)(.394)(1/18 + 1/825)}}$$

$$Z = 2.88$$

the extent to which the parole rate for each group differs from the white and black inmate parole rate than did those based on the unadjusted data.[98]

C. *The Binomial Trials are Independent of Each Other*

The third condition assumed to be satisfied in applying the binomial model requires that successive trials or drawings be independent of one another, in the sense that subsequent to any given number of selections there remain the same probabilities of selection for each of the two groups at the next selection (*i.e.*, the probabilities on each selection are not affected by the outcome of any other selection). This condition is satisfied when each selection is made truly "independently" in a physical sense and when the selections are not tied to each other in some way. This condition is violated when, for example, the method of selection is altered as the result of the outcomes on other trials or the unit of selection is groups of persons and not individuals (and the model is applied to data on the individuals selected).

An example where the independence condition was not satisfied is provided in the case of *Bryan v. Koch*.[99] In *Bryan*, the plaintiffs sought to prevent the City of New York from closing or reducing the number of beds in 4 of the 13 acute care municipal hospitals in the city on the ground that the selection of those four particular hospitals for closure or bed reduction was tainted by discrimination against blacks and Hispanics.[100]

At trial, the plaintiffs compared the proportion of minority patients or beds in the hospitals slated for closure or reduction with the proportion of minority patients or beds in the entire municipal hospital system.[101] Table 3 gives the numbers of minority and non-minority patients served in the 13 municipal hospitals based on a 1979, one-day census of emergency room patients.[102]

---

[98] Still another approach to statistical testing when discrimination against two or more groups is alleged is what is referred to as the chi-square test of homogeneity of binomial proportions. For a discussion of this test in a standard statistics text, see SNEDECOR & COCHRAN, *supra* note 4, at 201. The chi-square test is a test of whether or not the selection rates are the same for *all* groups. If the test at a given level of significance finds that there are differences, then a subsequent test is used to establish the significance of a difference for any particular group. If this test were to be applied to the parole data in *Nebraska Inmates*, black and white inmates would first be grouped to form one category, because discrimination against blacks was not alleged. If this were not done, a test result that indicated that there were statistically significant differences among the groups might be attributable, at least in part, to a difference between blacks and whites, a difference that is without interest in the case. Applied to the three categories of white-black combined, Native American, and Mexican-American, the test finds differences between the selection rates for the groups significant at the 1% level.

[99] 492 F. Supp. 212 (S.D.N.Y. 1980), *aff'd*, 627 F.2d 612 (2d Cir. 1980).

[100] 492 F. Supp. 212, 215-17.

[101] *Id.* at 218-21.

[102] *Id.*

Table 3

Distribution of Minority Patients
Among Municipal Hospitals

Bryan v. Koch

Hospitals	Minority Patients	Non-Minority Patients	Total
4 Hospitals Selected for Closure or Reduction	1111	162	1273
9 Hospitals Not Selected for Closure or Reduction	1760	548	2308
Total System (13 Hospitals)	2871	710	3581

Although the plaintiffs in *Bryan* did not carry out exactly this calculation, applying the binomial method they applied to other data on beds to these data on patients, plaintiffs would proceed as follows. Taking as "n" the total number of patients "selected" for closure or reduction, 1273, and as "p" the proportion of minority patients in the entire eligible population, .802 (= 2871/3581), the plaintiffs would compare the observed number of minority patients selected, 1111, with the expected number, 1021, calculate a Z-score equal to 6.33, and conclude that the difference was highly significant.[103]

As the district court found, however, the application of the binomial model was erroneous, because the model presumed that *patients* were independently selected for "closure or reduction" when in fact hospitals were selected for closure or reduction.[104] When a hospital is closed, all the patients in it are "closed" out, so that there is no process of independently selecting patients. The difference is crucial for the quantity of evidence pointing towards discriminatory selection. If patients are independent units, then there is very strong evidence that selection is in some manner related to minority status. But if hospitals are the units, then since only 4 of 13 hospitals were selected, it is much more probable that an outcome that picked two or even three or four

---

[103] The calculations for the binomial test are as follows:

(1) n = 1273
p = .802
o = 1111
(2) E = 1273 x .802 = 1020.95
(3) SD = $\sqrt{n \times p \times (1 - p)}$ = 14.22
(4) Z = (1111 - 1020.95)/14.22 = 6.33

The probability of observing a Z-score as large or larger than 6.33 is less than .001.

[104] 429 F. Supp. 212, 219-20.

"more minority" hospitals could occur by chance even if hospitals were picked for closure with the same independent probability.

Since the sample of four hospitals is a substantial proportion of the population of 13, the hypergeometric distribution is used to test the hypothesis that *hospitals* were selected for closure with the same probability for both "more minority" and "less minority" hospitals.[105] The table below displays the data for the 13 hospitals.

Table 4

Selection of Hospitals for "Closure or Reduction"

Bryan v. Koch

	"More Minority" Hospitals	"Less Minority" Hospitals	Total
Selected	2	2	4
Not Selected	4	5	9
Total Eligible	6	7	13

The probability, calculated under the hypergeometric distribution, that two or more hospitals selected for closure or reduction would be chosen among the "more minority" hospitals is .68. The value of .68 of observing the result, given nondiscriminatory selection, is actually greater than .5, and, of course, substantially exceeds the commonly cited thresholds of .01, .05, and .10. It is dramatically greater than the probability calculated under the assumption that patients were independently selected given above, which was less than .001. Thus, as this case illustrates, there can be serious miscalculations of the statistical significance of observed differences when the binomial model is applied to selections that are not truly independent.

IV. Miscellaneous Problems with the Binomial Model as Used in Discrimination Litigation

A. *Defining "p" as the Population Selection Rate*

One further misapplication of the binomial model that has occurred in discrimination litigation is that courts have sometimes mistakenly defined "p" —

---

[105] A "more minority" hospital is defined as one having a percentage of minority patients greater than the median percentage of minority patients for the municipal hospital system as a whole. (This median was 77.8%.) A "less minority" hospital is one whose minority percentage is less than or equal to the median. Similar results are obtained when other possible definitions of a "more minority" hospital are chosen.

An alternative test to the hypergeometric for a somewhat different measure of discrimination that was also presented at trial in this case is the Mann-Whitney rank sum test. This test examines the hypothesis that the hospitals selected for closure are a random sample of the entire group of hospitals. Hospitals are ranked by percent minority, and the test asks if the sum of

the probability of success on each binomial trial — not as the proportion of the eligible population who are members of the disfavored group but rather as the proportion of the population as a whole (or of the favored group) who succeeded in being selected.[106] These latter proportions, or selection rates, which do figure in the difference-in-proportions test and could be analyzed by that test, play no explicit role in the binomial model. Using one or the other of them as "p" in the binomial equations leads to error.

For example, consider a situation in which 100 persons, 60 males and 40 females, have been hired from a pool of 2,000 applicants consisting of 1,000 males and 1,000 females. In applying the binomial model to test whether the underrepresentation of females among the hirees is statistically significant, we would define "p" as the proportion of females in the applicant pool, which is 1000/2000, or .5, and "n" as the number of persons hired, which is 100. The binomial calculations would proceed as follows:

$$E = n \times p = 100 \times .5 = 50$$
$$O = 40$$
$$SD = \sqrt{n \times p \times (1-p)} = \sqrt{100 \times .5 \times .5} = 5$$
$$Z = \frac{E - O}{SD} = \frac{50 - 40}{5} = 2.0$$

A Z-score of 2.0 means that the female shortfall is significant at both the 5 percent level and (though at the borderline) under the *Castaneda* two- to three-standard-deviation criterion.

On the other hand, if "p" is defined in terms of selection rates, the results are quite different. In this example the overall selection rate is 100/2000 or .05, and the male selection rate is 60/1000 or .06. The logic of this mode of analysis is that females would have been selected at one or the other of these rates absent discrimination; therefore, these rates are multiplied not by the number of selections but by the number of females in the eligible pool, which in this example is 1,000 (*i.e.*, n = 1,000), to obtain an expected number of female hires. The binomial calculations would then proceed as follows with the redefined terms "n" and "p" designated "n*" and "p*").

p = the overall selection rate*

$$E = n^* \times p^* = 1000 \times .05 = 50$$[107]
$$O = 40$$

---

the ranks of those selected for closure is significantly different from the same statistic for those not selected for closure. *See* F. MOSTELLER & R. ROURKE, STURDY STATISTICS: NONPARAMETRICS AND ORDER STATISTICS 56 (1973).

[106] *See* Rivera v. Wichita Falls, 665 F.2d 531 (5th Cir. 1981) (alleged discrimination against Mexican-Americans in testing of police recruits: "p" defined as failure rate for white applicants); Hameed v. Iron Workers, 637 F.2d 506 (8th Cir. 1980) (alleged discrimination against blacks in admittance to union apprenticeship program: "p" defined as selection rate for all applicants).

[107] It should be noted that this expected number of 50 equals that calculated above using the proper binomial model. That this is a general result can be seen from the algebraic identity:

$$SD = \sqrt{n^* \times p^* \times (1\text{-}p^*)} = \sqrt{1000 \times .05 \times .95} = 6.89$$

$$Z = \frac{E - O}{SD} = \frac{50\text{-}40}{6.89} = 1.45$$

p^* = *the male selection rate*

$$E = n^* \times p^* = 1000 \times .06 = 60$$

$$O = 40$$

$$SD = \sqrt{n^* \times p^* \times (1\text{-}p^*)} = \sqrt{1000 \times .06 \times .94} = 7.51$$

$$Z = \frac{E - O}{SD} = \frac{60 - 40}{7.51} = 2.66$$

The first Z-score of 1.45 is not significant at the 5 percent level, and the statistical evidence of female underrepresentation among persons hired would probably be rejected by most courts as not probative of discrimination. The second Z-score of 2.66 is significant at the 1 percent level and would probably be accepted by many courts as highly probative of discrimination against females in this hiring process. Both results are wrong, however, understating in one case and overstating in the other the actual statistical significance of the female shortfall among the persons hired, which is accurately estimated by the binomial model only if "p" is defined in terms of the proportion of females in the eligible pool.

The problem with defining "p" in the binomial model in terms of selection rates is that doing so violates the binomial requirement that the probability of success on each trial be fixed at the beginning of the binomial experiment and not vary as a function of the number of trials in the experiment or the outcomes of those trials. The binomial model in effect compares this known quantity (*e.g.* the gender composition of the applicant pool) against the results of the selection process (*e.g.*, the gender composition of the population of hirees). Selection rates, on the other hand, are themselves derived from the results of the experiment. And while racial or gender disparities in selection rates are meaningful indicators of possible discrimination against the adversely affected groups, the statistical significance of such disparities must be determined through different statistical tests — principally the difference-in-proportions test — and not the binomial.[108]

$$\text{(number of selections)} \times \frac{\text{(number of females in eligible pool)}}{\text{(number of persons in eligible pool)}}$$

$$= \frac{\text{(number of selections)}}{\text{(number of persons in eligible pool)}} \times \text{(number of females in eligible pool)}$$

The product on the left-hand side is simply "n × p" of the proper binomial model and that on the right-hand side is the "$n^* \times p^*$" described above when the probability of success on each binomial trial is defined as the overall selection rate. Thus, the problem with the latter approach comes not in determining the expected results of the selection process but in calculating the standard deviation, which is used to gauge the significance of the difference between the expected and observed results.

[108] *See supra* notes 53-58 and accompanying text for a discussion of the difference-in-

B. *Small Size of the Sample Selected*

In *Castaneda*, the Supreme Court recognized that the binomial analysis it was using was valid only for "large samples."[109] The sample to which the Court was referring was the total number of persons called to serve as grand jurors in Hidalgo County during the period it was analyzing.[110] And, in general, for the selection situations to which the binomial is applied in discrimination litigation, the relevant "sample size" is simply the total number of selections — the "n" of the binomial equations.

Expected values and standard deviations for the binomial distribution are calculated in the same fashion for small samples as for large samples. But for small samples the tests of statistical significance described in this article and in *Castaneda* no longer apply because those tests rely on the normal approximation to the binomial, and the binomial distribution approximates the normal distribution well only for sample sizes of sufficient magnitude. Accordingly, for small samples the Z-statistic obtained by dividing the difference between the expected and observed results of the binomial experiment by its standard deviation does *not* have a standard normal distribution. The statistical significance of any such difference, therefore, cannot be determined accurately by simply looking up the value of the Z-statistic on a standard normal probability table. For the same reason the *Castaneda* Court's suggestion that differences between the expected value and observed results of more than two to three standard deviations are statistically significant[111] is also not reliable when dealing with small samples.

The sample size required to make use of the normal approximation to the binomial is a function of "p" — the probability of success on each binomial trial. The more extreme "p" is, *i.e.*, as it gets closer to either zero or to 1.0, the larger the sample size must be before the binomial distribution will be accurately approximated by the normal distribution.[112] As a rough rule of thumb, if

---

proportions test.

[109] 430 U.S. at 496 n.17.

[110] *Id.* at 487.

[111] *Id.* at 496 n.17.

[112] For small sample sizes and extreme values of "p," the binomial distribution is very skewed, with the expected value and the other most probable outcomes bunched closely at one end of the distribution. For example, for n = 10 and p = .9, the exact binomial distribution is as follows:

Number of Successes	Probability
10	.349
9	.387
8	.194
7	.057
6	.011
5	.001

(For each outcome of less than 5 successes the probability is less than .001.) The normal distribution, on the other hand, is symmetric about the expected value with probabilities declining gradually as you move away from the expected value in either direction. A normal curve cannot be fitted very well to a highly skewed distribution, such as the binomial with n = 10 and p = .9. In

the expected value — which equals the product of "n" and "p", — is at least three standard deviations from the extremes of the distribution (*i.e.*, from zero or "n") the normal approximation to the binomial will be quite accurate.[113] When this condition is not satisfied, exact binomial probabilities for the observed outcomes may be calculated.[114]

CONCLUSION

Statistical evidence has taken on an increasingly important role in litigation involving claims of discrimination against particular classes of people. The Supreme Court has recognized the utility of formal statistical tests in assessing the probative value of such evidence in cases alleging discrimination in violation of the Constitution and Title VII of the Civil Rights Act of 1964. The particular statistical model used by the Court — the binomial model — is applicable to many types of selection processes that are frequently the subject of dispute in discrimination litigation, such as jury selection and various employment practices. In those cases, the binomial model is a powerful and, because of its relative simplicity in conception and application, particularly useful analytic tool for determining whether observed disparities among different groups in the results of a selection process support an inference of discrimination in that process. Lower courts, following the Supreme Court's lead, have frequently made good use of the binomial model in evaluating the often conflicting statistical evidence presented by parties in discrimination litigation.

When any statistical model is applied to a set of data, however, care must be taken to ensure that the requirements of the model are met in the process that generated the data or, if not met, that the divergence from those requirements does not invalidate the conclusions that are to be drawn from the model. Lower courts have not always been careful when applying the binomial model to be sure the requirements of the model — which were not explicitly spelled out by the Supreme Court — have been satisfied. The applicability of the binomial model does depend on a number of assumptions about the selection process being studied. When these assumptions are not met, the binomial model may not be used or may only be used if certain adjustments are made.

---

such cases probability estimates for binomial outcomes derived from a normal distribution may be inaccurate.

[113] In these circumstances it appears that the maximum error in estimating the probability that a difference as great as that observed could be the result of random fluctuations is no more than .025. MOSTELLER, ROURKE & THOMAS, *supra* note 4, at 290. This is an estimate of the upper limit of the potential error, and in practice the normal approximation is generally more accurate, not only when the expected value is further from the extremes than three standard deviations, but in many cases when it is within three standard deviations of the extremes.

[114] For small sample sizes, these calculations are feasible although cumbersome. *See supra* note 35 and accompanying text. Of course, computers can perform such calculations rapidly and accurately. Furthermore, there are binomial probability tables available, which for some sample sizes have been reproduced in readily available statistics texts. *See, e.g.*, MOSTELLER, ROURKE & THOMAS, *supra* note 4, at 475-91 (binomial tables for sample sizes 2 to 25 and for 13 values of "p" from .01 to .99).

This article has analyzed the requirements imposed by those assumptions, described selection situations where those assumptions are not met, and suggested alternatives or adjustments to the binomial that may be used in those circumstances. By insisting on both a careful examination of the assumptions underlying the binomial model before it is applied as a test for discrimination and the use of the alternatives and adjustments to the model discussed in the article when those assumptions are not met, courts and litigants in discrimination litigation will have greater assurance that the inferences they draw from the results of statistical tests are ones the evidence truly supports.

BOSTON COLLEGE LAW REVIEW

VOLUME XXIV JULY 1983 NUMBER 4