Statistics and the Law: Hypothesis Testing and Its Application to Title VII Cases

Louis J. Braun
Statistics and the Law: Hypothesis Testing and Its Application to Title VII Cases

By Louis J. Braun*

Statistics form a useful tool in the determination of whether defendants in a Title VII action have engaged in employment discrimination. Statistics also can determine whether disparities in the form of an underrepresentation of minority group members in particular jobs or differences in pay, benefits, or conditions of em-

* Associate Professor of Mathematics and Statistics, Pace University, New York, New York. B.A., 1963, New York University; Ph.D., 1971, City University of New York.

The author wishes to thank Donald H. Zeigler and Donald L. Doernberg, Associate Professors of Law, Pace University School of Law, for their helpful suggestions in the preparation of this Article.

1. Statistics may be defined loosely as the collection, correlation, and interpretation of data. For a detailed explanation of statistical analysis see, D. Huntsberger, D. Croft & P. Billingsley, Statistical Inference for Management and Economics (2d ed. 1980) [hereinafter cited as Huntsberger].

2. Title VII of the Civil Rights Act of 1964, as amended, 42 U.S.C. §§ 2000e to 2000e-17 (1976 & Supp. II 1978), makes it unlawful 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 . . . .” Id. § 2000e-2(a)(i). In addition to governmental agencies and departments, private parties have causes of action under Title VII that allow them to recover damages for injuries sustained as a result of discriminatory treatment directed against them as employees or potential employees. See Johnson v. Goodyear Tire & Rubber Co., 491 F.2d 1384 (5th Cir. 1974).

ployment can be attributed solely to chance factors. Plaintiffs often have relied on statistical analysis to meet the initial burden of establishing a prima facie case of discrimination.\(^4\) Defendants too have used statistics to show, among other things, that the employment practices which are alleged to be discriminatory amounted to bona fide job requirements\(^5\) or to rebut the statistical evidence offered by plaintiffs.\(^6\) In *International Brotherhood of Teamsters v. United States*,\(^7\) the Supreme Court recognized the importance of statistical analysis in cases in which the existence of discrimination is a disputed issue.\(^8\) Furthermore, statistics showing an extraordinarily small number of minority employees have been held to establish a violation of Title VII as a matter of law.\(^9\)

Some courts nevertheless have shown a reluctance to use statistics in employment discrimination cases, either because they did not understand the analysis, found the theory too complicated,\(^{10}\) or

---

4. In *McDonnell Douglas Corp. v. Green*, 411 U.S. 792, 802 (1973), the Court held that in a private Title VII action, the plaintiff will satisfy the burden of establishing a prima facie case of discrimination if he or she shows: "(i) that he belongs to a racial minority; (ii) that he applied and was qualified for a job for which the employer was seeking applicants; (iii) that, despite his qualifications, he was rejected; and (iv) that, after his rejection, the position remained open and the employer continued to seek applicants from persons of complainant's qualifications." In subsequent cases, courts have held that the plaintiff need not have applied for the position if he or she can show that, given the defendant's conduct, such an act would be futile. See, e.g., *Rodriguez v. East Texas Motor Freight Sys., Inc.*, 505 F.2d 40, 55 (5th Cir. 1974), *vacated on other grounds*, 431 U.S. 395 (1977).

In such "disparate treatment" cases, proof of the defendant's discriminatory motive is critical although it may be inferred from the mere fact of differences in treatment. See *International Bhd. of Teamsters v. United States*, 431 U.S. 324, 355 n.15 (1977). However, courts also have held that prima facie cases of employment discrimination can be established, by both private and governmental plaintiffs, by showing that the defendant's policies had a "disparate impact" on minorities without having to prove that the defendant intended to discriminate. *Id.* Disparate impact analysis focuses on the consequences of the defendant's employment practices, not the motives behind them.


8. *Id.* at 339.


because the results were not grounded in absolute certainty.\textsuperscript{11} Even those courts accepting statistical analysis, but forced to rely on the interpretations and understandings of expert witnesses, sometimes express a wariness and discomfort in using statistics to resolve legal issues.\textsuperscript{12} In \textit{International Brotherhood of Teamsters}, for example, the Court stated: “We caution only that statistics are not irrefutable; they come in infinite variety and, like any other kind of evidence, they may be rebutted. In short, their usefulness depends on all of the surrounding facts and circumstances.”\textsuperscript{13} Although statistics can be misused in developing an aura of support for a position not supportable by the evidence,\textsuperscript{14} rather than bar or severely curtail its use, steps can and should be taken to impose restrictions which would disallow its improper use only. If the rudiments of statistical analysis were understood by the courts, they would have a better perception of what constitutes its proper use and could exercise a more constructive control over its admissibility as evidence.

This Article discusses a limited but central area of statistical analysis applicable to Title VII cases.\textsuperscript{15} Although the techniques and theories developed herein all have been used by expert witnesses in such cases, their synthesis into a cohesive theory of evidence should foster a greater understanding and thus a more productive use of statistics in employment discrimination litigation.

The focus of the Article will be on the field of statistical analysis known as \textit{hypothesis testing},\textsuperscript{16} with an initial discussion of the problems in determining the proper population of available em-

\textsuperscript{11} See, e.g., Chance v. Board of Examiners, 458 F.2d 1167, 1173 (2d Cir. 1972).

\textsuperscript{12} A large measure of the distrust with which courts often view statistics stems from the popular reputation of statistics as being an analytical device that can be used to twist the truth and prove whatever it is one wants to prove. See, e.g., D. Huff, How \textit{To Lie With Statistics} (1954).

\textsuperscript{13} 431 U.S. at 340.

\textsuperscript{14} See Huntsberger, \textit{supra} note 1, at 4-8.

\textsuperscript{15} This Article is not intended to offer a complete and detailed account of each and every statistical technique that can be used in a Title VII action, but rather to present and explain the most useful of the analytical techniques which apply to such cases. Thus far, the major cases in employment discrimination which turned on statistical evidence could all have been analyzed using the techniques contained herein. Nor are the techniques developed in this Article restricted to use in Title VII actions only. They can be used in such disparate fields as antitrust litigation, environmental law, drug effectiveness cases, and disputed election actions.

ployees or "qualified labor pool" to be used in the analysis. The Article will develop methodology and mathematical techniques relevant to the determination of whether a proper proportion\textsuperscript{17} of a certain minority is represented in the employ of a given company, or whether the differences between the proportions or mean\textsuperscript{18} salaries, increments, or tenure without promotion of various groups of people employed in given jobs are significant enough to warrant an inference of discrimination. Rather detailed but fundamental mathematical analysis is presented throughout this Article which should aid in fostering an understanding helpful both to potential litigants and to the courts in their use of statistics.

Populations

Two methods by which a plaintiff in a Title VII action may use statistical analysis to establish a prima facie case of discrimination in employment are: (1) demonstrating that the proportion of minority employees in a defendant company working in particular jobs is significantly smaller than the proportion of minority members in the general population able to do those jobs\textsuperscript{19} and (2) showing that a defendant employer's hiring or promotion criteria discriminate against a given minority by excluding or disqualifying a greater proportion of its members than other groups within the population of qualified workers.\textsuperscript{20} A defendant attempting to rebut such arguments and evidence may challenge the plaintiff's basis for establishing the population of available employees.\textsuperscript{21} Defining the

\textsuperscript{17} A "proportion" is a ratio, \(x/n\), where \(x\) is the number of observations describing a given event, and \(n\) is the total number of observations. A proportion also can be given as a decimal or a percentage. For example, if there are 100 women in a work force of 2,000 the proportion of women in that work force is: \(100/2,000 = 0.05 = 5\%\).

\textsuperscript{18} A "mean" is an average. It is computed by adding the given arithmetic observations and then dividing by the total number of such observations. For example, if \$5.65, \$5.93, \$6.07, and \$5.75 were the hourly rates of pay for the four workers on a given assembly line, the mean hourly rate would be:

\[
\frac{(5.65 + 5.93 + 6.07 + 5.75)}{4} = \frac{23.40}{4} = 5.85.
\]

\textsuperscript{19} See, e.g., EEOC v. Local 14, Int'l Union of Operating Eng'rs, 553 F.2d 251 (2d Cir. 1977); Wetzel v. Liberty Mut. Ins. Co., 508 F.2d 239 (3d Cir.), cert. denied, 421 U.S. 1011 (1975).


relevant population thus can be central to the issue of discrimina-
tion and is often a matter of contention between the parties. In
Hazelwood School District v. United States, for example, the
plaintiff successfully had established a prima facie case of employ-
ment discrimination by demonstrating that there was a significant
disparity between the proportions of black teachers in the Hazel-
wood school system and the population of teachers in the St. Louis
metropolitan area in which Hazelwood is located. The defendant
successfully argued before the Supreme Court that the city of St.
Louis should possibly be excised from this population which would
result in the elimination of a large portion of black teachers from
consideration and bring the relevant proportions generally into
proper alignment. The Court therefore vacated the appellate court
decision, remanding the case to the trial court for a determination
of the appropriate comparative population.

The two major criteria used in determining the population of
available employees for a specific job have been the job qualifica-
tions including the skills, talents, or abilities needed to perform the
job and the geographic area from which the employees are drawn.
Absent a showing of special job qualifications, the population is
determined by considering only the appropriate geographic area
and including all persons in the employable age bracket living
within that area. If, however, the jobs in question could be filled
only by highly trained individuals or persons with special talents
or abilities, the dominant criterion then would be the job qualifica-
tions, and a geographic limitation may not even be considered.
In these circumstances, the appropriateness of job qualifications
must be addressed on a case by case basis and often is a central
issue in the action.

The determination of the appropriate geographic area to be
used in defining the relevant population also has been approached
on a case by case basis. Courts have used a variety of measures in

23. Id. at 310-13.
25. See, e.g., Hester v. Southern Ry. Co., 497 F.2d 1374, 1379 n.6 (5th Cir. 1974); Spur-
lock v. United Airlines, Inc., 475 F.2d 216 (10th Cir. 1973).
defining the relevant geographic areas, including political subdivisions, metropolitan areas, and the jurisdictions or areas served by the defendants. To avoid the potential inconsistencies in such an ad hoc approach, the Equal Employment Opportunity Commission (EEOC) has designated the “Standard Metropolitan Statistical Area” (SMSA) as the geographic area to be used in Title VII actions. The SMSA may be a useful geographic parameter if the locus of employment is at or near its center; however, its usefulness diminishes as this locus is moved closer to its edges. A defendant company on a boundary of an SMSA will not draw employees from the distant parts of the region except under the rarest of circumstances. Conversely, the company probably would draw part of its workforce from neighboring towns and counties that are not a part of the SMSA. Although many courts now are adopting the SMSA as the geographic area of the population, its acceptance is by no


31. SMSAs are geographic units developed by the Office of Management and the Budget (OMB) comprising cities, towns, and counties in and around major United States cities.

32. For example, the New York City SMSA includes Westchester and Putnam Counties, New York in the north and Brooklyn, New York in the south. For an employee living in Brooklyn to commute the 60 plus miles by auto to a plant in northern Westchester or Putnam during the rush hours would take at least two and one-half hours one way, assuming only normal delays. If the employee was at attempt to commute by rail, he or she would first have to get to a subway station (usually by bus), take the subway to Grand Central Station (which may involve transferring trains), take the railroad train to the Westchester or Putnam stop, and finally take a taxi to the place of employment. This trip easily could approach three hours in duration even assuming no delays and efficient connections. Conversely, a worker living in the non-SMSA counties of Orange or Sullivan, New York would be within a reasonable commuting distance from a worksite in northern Westchester or Putnam.

33. In the rare instances where the place of employment is not located within an SMSA, the EEOC uses the “Economic Area” (EA), which is a geographic unit similarly defined along county lines. See The BEA Economic Areas: Structural Changes and Growth 1950-1973, Survey of Current Business (Nov. 11, 1975).
means universal,\textsuperscript{34} nor is its accuracy unassailable.

The problem inherent in the use of SMSAs was demonstrated in \textit{Abron v. Black \& Decker Mfg. Co.}\textsuperscript{35} The court in \textit{Abron} rejected the defendant's contention that the geographic area to be used in defining the relevant population should be the area within a twenty mile radius of the plant in favor of the SMSA of Baltimore, Maryland.\textsuperscript{36} The court concluded that the defendant had engaged in employment discrimination despite substantial evidence that the SMSA did not accurately reflect the geographic area in which the Black \& Decker workforce resided.\textsuperscript{37}

Although exclusive reliance on the SMSA to define the geographic area of the relevant population when dealing with companies located in or near major United States cities may lead to inequitable results, continuing to determine these areas on an ad hoc basis easily could lead to inconsistent results on similar facts. For example, had the court in \textit{Abron} accepted the defense contention that the appropriate geographic area should be defined as that within a twenty mile radius of the Black \& Decker plant, it probably would have found no case of employment discrimination had been established.

A uniform approach for determining the geographic area of the population of available employees that avoids inflexible standards which may be inapplicable to the facts of an individual case, as demonstrated by the use of the SMSA, yet offers predictability of result and consistency of decision may be established by the use of a single mathematical formula. To be accepted by the courts, the formula would have to be simple to apply and reflect adequately the relative mobility of the workforce in the specific regions in question.

Absent natural geographic criteria,\textsuperscript{38} the dominant factor to be

\textsuperscript{34} See cases cited in notes 28-29 \textit{supra}.
\textsuperscript{36} \textit{Id.} at 1105.
\textsuperscript{37} \textit{Id.} This evidence indicated that: (1) 85\% of the Black and Decker workforce resided within a twenty mile radius of the plant; (2) no direct public transportation existed between major intersections in Baltimore and the plant (25 to 30 miles in distance); and (3) blacks in the region were less likely to own cars than whites.
\textsuperscript{38} Such criteria would include state or municipal requirements of residence as a condition of employment. Also, if the wages paid in one state were significantly higher than the wages paid in an adjacent state for comparable work done, the common state boundary would act as a natural geographic boundary for the labor pool availability area. See, \textit{e.g.}, United States v. County of Fairfax, 19 FEP Cases 753 (E.D. Va. 1979).
considered by such a formula is the distance traveled by the workforce to the place of employment measured in terms of either miles or time.\textsuperscript{39} The time, effort, and expense of a commute is generally a major consideration of a potential employee in deciding whether to apply for or accept a particular job.\textsuperscript{40} It would not be much of an improvement over the SMSA, however, to draw a circle of a given radius centered at the place of employment and label that region the geographic area. Such a construct is too rigid, eliminating from consideration those potential employees who would be willing to commute to work from locations outside of the constructed area; it also fails to consider that the highway network or mass transit facilities in the region may make it faster and easier to commute to the worksite from specific locations outside the circle than from some locations within it.\textsuperscript{41}

Taking into account all of the above considerations, the geographic area of the population of available employees should be the aggregate or union of every city, town, and county in which a given percentage of the populace resides within a specified distance from the place of employment.\textsuperscript{42} The demographic information is readily available, the construction is not difficult, and it allows for uniformity without sacrificing adequate consideration of relevant individual factors. The construct resembles that of an SMSA except that the focus or “practical center” is the place of employment, not a major city.\textsuperscript{43} By requiring only a percentage of the populace of a city, town, or county to reside within a given

\textsuperscript{39} In urban areas, traffic congestion and mass transit facilities affect the correlation between miles traveled and time spent in transit. In these locales, it is more appropriate to measure commuting distances in terms of time. However, the cost of commuting, either by auto or rail, cannot be ignored because it generally is determined outside central cities according to the miles traveled. In rural areas where traffic congestion and mass transit are not major factors influencing commuting, the classical measure of distance in miles should serve well.


\textsuperscript{41} For example, in the New York metropolitan area, it takes 45 minutes to commute from White Plains to midtown Manhattan, a distance of more than 20 miles, while it takes over one hour to commute from the Canarsie section of Brooklyn to midtown, a distance of less than 15 miles.

\textsuperscript{42} A reasonable choice would be the union of all political subdivisions where more than two-thirds of the populace live within a one hour and a 40-mile commute to the place of employment, thus taking into account both distance and time. Whatever measures finally are adopted, however, should be accepted uniformly so as to promote consistency and order.

\textsuperscript{43} Thus, under this construction, a worksite in Manhattan would be expected to draw employees from Brooklyn, while one in Putnam County would not. See note 32 supra.
commuting distance from the worksite, consideration of potential employees who would be willing to commute from locations outside the region is to a large degree retained. 44 This formula meets the requirements of ease of construction and reasonableness, and if adopted, should aid courts in determining the geographic area of the population with a consistency previously lacking.

It should be noted, however, that this construction is not appropriate in all cases. When the personnel needed are required to possess such unique skills or talents that an employer is forced to recruit outside such areas, 45 when the job itself entails moving about with no central starting point for the employees, 46 or when a town, city, or county within the constructed region has instituted a program which would bias the results if included, 47 an alternate construction would have to be considered. Under such circumstances an alternative generally would be apparent. 48

**Statistical Foundation**

Having defined a population of available employees, a comparison can be made between that population and the defendant's workforce, or among distinct groups within the defendant's workforce, to determine whether any existing disparity in the treatment of minority and nonminority workers is significant enough to warrant an inference of discrimination. The law initially presumes that no such disparity exists, and that any such difference is not significant enough to create an inference of discrimination. To counter this presumption and establish a prima facie case, a plaintiff must first demonstrate that a disparity of treatment ex-

44. For example, if only 70% of the populace of Brooklyn resides within a specified commuting distance to a given place of employment, all qualified individuals who live in Brooklyn would be considered a part of the population of available employees, including those who do not live within the specified commuting distance.


ists. If such a disparity is found, use of the statistical technique of “hypothesis testing”\(^49\) can determine whether this difference in treatment is the result of discriminatory acts by the defendant.

The Null Hypothesis

When the law presumes that there is no significant difference between two or more statistical data groupings, as opposed to presuming that such a difference does exist, it is formulating a null hypothesis.\(^50\) A null hypothesis postulates equality rather than inequality, presuming that whatever statistically demonstrated inequality does exist is due to chance factors.\(^51\) It is mathematically more straightforward to determine whether a null hypothesis should be rejected as false by testing the degree of disparity against the expected sampling error, than to determine the falsity of a hypothesis that is not null.\(^52\) Placing the burden on the plaintiff in a Title VII action to establish a prima facie case is functionally equivalent to assuming that there is no difference in the treatment by the defendant of specified groups in its workforce—a null hypothesis.

The alternate hypothesis to a null hypothesis postulates a difference between two or more statistical data groupings and may take several forms. Given a null hypothesis that no difference exists between two arithmetic observations \(x\) and \(y\) \((x = y)\), its negative could be: (1) the first observation is greater than the second \((x > y)\); (2) the first observation is less than the second \((x < y)\); or (3) the first observation is different from the second \((x \neq y)\).

The formulation of an alternate hypothesis is important when determining whether a defendant discriminated in violation of Title VII inasmuch as rejection of a null hypothesis is statistically equivalent to accepting a constructed alternative. Similarly, acceptance of the null hypothesis automatically means rejection of any constructed alternative.\(^53\) For example, if the first observation \((x)\) is the proportion of blacks in the defendant’s workforce, and the second observation \((y)\) is the proportion of blacks in the popu-

\(^49\) See note 16 supra.

\(^50\) For a more detailed description of the null hypothesis, see generally Freund & Williams, supra note 16, at 281; Huntsberger, supra note 1, at 274-76.

\(^51\) E. Babbe, Survey Research Methods 309 (1973).

\(^52\) See Freund & Williams, supra note 16, at 280-281.

\(^53\) Id. at 275-76, 283.
lation of available employees, the three possible alternate hypothe-
ses would be: (1) the proportion of blacks in the defendant's
workforce is greater than the proportion of blacks in the popula-
tion of available employees \(x > y\); (2) the proportion of blacks in
the workforce is less than the proportion of blacks in the popula-
tion \(x < y\); and (3) the proportion of blacks in the workforce is
different from that in the population \(x \neq y\). Typically in Title
VII cases the court is concerned with the possibility of inadequate
minority representation in the workforce, focusing on alternate hy-
pothesis (2) rather than (1) or (3). In such cases, a rejection of
the null hypothesis—that the proportion of blacks in the popula-
tion is generally reflected by the proportion of blacks in the
workforce—would be equivalent to accepting the proposition that
the proportion of blacks in the workforce is significantly smaller.
In a case of alleged reverse discrimination, however, the alternate
hypothesis would be, for example, that the proportion of blacks in
the workforce is larger than that in the population.

Type I Error

Accepting or rejecting any hypothesis entails the risk of error.
Rejecting a hypothesis when it is in fact true is classified as a Type
I error. Demanding that a given hypothetical statement be ac-
cepted unless the weight of the evidence exceeds a selected thresh-
old limits the possibility of making such an error. The possibility
remains that a false hypothetical statement will be accepted as
true because the weight of the evidence presented falls below the
selected threshold. Because the presumption in Title VII cases is
that no discrimination exists (a null hypothesis) with the initial
burden on the plaintiff to disprove this proposition by a prepon-
derance of the evidence (a selected threshold), the courts appear
more reluctant to blame an employer who is actually not guilty of
discrimination (rejecting a true null hypothesis and making a Type
I error with respect to the innocence of the employer defendant)

54. See, e.g., Hazelwood School Dist. v. United States, 433 U.S. 299 (1977);
1978), rev'd, 608 F.2d 671 (6th Cir. 1979).
56. See Freund & Williams, supra note 16, at 276; Huntsberger, supra note 1, at
277-78.
57. Accepting a false hypothesis is a Type II error. See Freund & Williams, supra
note 16, at 276; Huntsberger, supra note 1, at 277-78.
than to exculpate a defendant who did discriminate unjustly.\textsuperscript{58}

**Level of Significance**

The question arises as to the degree of proof courts should require to reject a null hypothesis and hold that a plaintiff has established a prima facie case of discrimination. Put another way, what is the probability of error that courts should tolerate before finding that discrimination exists? This tolerable possibility of making a Type I error is called the *level of significance*,\textsuperscript{59} and is denoted by the Greek letter alpha (\(\alpha\)). The choice of an \(\alpha\) value is one to be made by the courts based on what they feel is an adequate threshold for the burden of proof by a preponderance of the evidence.

The EEOC and some courts have settled on a level of 0.05.\textsuperscript{60} By this measure, the null hypothesis that a defendant did not discriminate would be rejected in favor of the alternate hypothesis of discrimination if there is less than a 5\% (0.05) chance that the rejection is an error. Thus, the plaintiff must establish that there is at least a 95\% chance that the defendant’s actions were discriminatory before the court will hold that a prima facie case has been established by a preponderance of the evidence.

**The Central Limit Theorem\textsuperscript{61}**

The techniques used to determine whether a null hypothesis should be rejected, given a certain level of significance, are predicated on the following statistical theory.\textsuperscript{62} Given a large number\textsuperscript{63} of *random samples*\textsuperscript{64} of equal and sufficient size\textsuperscript{65} taken from a

---

\textsuperscript{58} This parallels the presumption in criminal cases that a defendant is innocent until proven guilty.

\textsuperscript{59} See Freund & Williams, *supra* note 16, at 283; Huntsberger, *supra* note 1, at 278.


\textsuperscript{63} This number is postulated to be large enough to accomplish the objectives to be enunciated later in this discussion. Such a number exists in all cases to be considered.

\textsuperscript{64} A random sample is formed by selecting a given number of elements of the population by procedures which ensure that no bias will ensue and that every distinct sample of the given size has an equal chance of being selected. To construct a sample which purposefully includes or excludes elements from given groups within the population, such as mandating that 50\% of the sample taken be comprised of women or limiting the elements of the sample to only a part of the geographical area of the population, creates a bias. Any results
given population, a statistical parameter such as a mean, a proportion, or a standard deviation is extracted from each sample. These extracted bits of statistical data, when taken together, always are going to be "normally distributed," and their mean will be the statistical parameter of the given population. In addition, the standard deviation of these extracted bits of data, called the standard error of the relevant parameter, is given by means of its own formula.

regarding characteristics of the population inferred from the data extracted from such a sample are tainted. Some courts have been forced to grapple with the question of randomness of a sample and disregard testimony grounded in one found to be biased. See, e.g., United States v. Georgia Power Co., 474 F.2d 906, 916 (5th Cir. 1973); Queen v. Dresser Indus., Inc., 456 F. Supp. 257, 264-65 (D. Md. 1978).

Some courts have demonstrated a reluctance to admit evidence proffered from samples which were small in size. See, e.g., Friend v. Leidinger, 446 F. Supp. 361, 366-67 (E.D. Va. 1977); aff'd, 588 F.2d 61 (4th Cir. 1978); Dendy v. Washington Hosp. Center, 431 F. Supp. 873, 876 (D.D.C. 1977); Johnson v. Shreveport Garment Co., 422 F. Supp. 526, 539-40 (W.D. La. 1976), aff'd, 577 F.2d 1132 (5th Cir. 1978). See also Dendy v. Washington Hosp. Center, 581 F.2d 990 (D.C. Cir. 1978). In remanding Dendy to the trial court, the court of appeals stated: "While the testimony of plaintiff's statistician was uncontroverted, the court held that it represents too slender a reed on which to rest the weighty remedy of preliminary relief. . . . While the numbers involved may have appeared small, statistical analysis showed them to reflect a discriminatory impact that could not reasonably be ascribed to chance alone." Id. at 992 (citations omitted). See also Chicano Police Officer's Ass'n v. Stover, 526 F.2d 431, 439 (10th Cir. 1975), vacated on other grounds, 426 U.S. 944 (1976), "(The smallness of the sample should not be grounds here for rejecting the proof. If it were, the tendency would be to deny employees in small plants the type of protection the civil rights statutes afford)."


A standard deviation measures the dispersion of a set of data. Its relative size depends upon how widely the data is dispersed. If the data are closely bunched, the standard deviation is small; if the data are scattered, it would be fairly large. See Freund & Williams, supra note 16, at 39, 45-46; Huntsberger, supra note 1, at 46-51, 63-65.

A normal distribution can be viewed as the distribution of observations of repeated measurements of a given physical phenomenon. The scores in a normal distribution congregate around the mean and taper off symmetrically as one moves further away from the mean in either direction. See Freund & Williams, supra note 16, at 224-25. An example would be the distribution of the observed number of heads obtained by repeatedly flipping a coin in the air 100 times. The first series of flips might produce 55 heads, the next series might produce 47 heads, the next 44 heads, etc. . . . The numbers 55, 47, 44, . . . would be normally distributed. If the phenomenon in question either rarely or almost always occurs, the distribution of observations of repeated measurements is not normal. See id at 189-91; Huntsberger, supra note 1, at 154-56 (a discussion of the Poisson distribution).

See Freund & Williams, supra note 16, at 250, 292, 315, 332 (problem 5).
Illustration 1: Assume that 250 employable adults are selected randomly from a population and given a certain examination, which 200 (80%) of them pass. Now assume that another 250 employable adults are chosen at random and given the same examination and 185 (74%) pass. Continuing this process of choosing 250 employable adults, testing them, and recording the proportion that pass, a large number of times, say 100,000, one would have 100,000 proportions extracted from the individual samples; 80% or 0.80 from the first, 74% or 0.74 from the second, and so forth. These 100,000 proportions would be normally distributed with their mean:

\[ p = \frac{0.80 + 0.74 + \ldots}{100,000} \]

being the actual proportion of adults in the population who would have passed the examination had it been administered to each and every one of them.\(^7^1\) Had another parameter such as the mean test score of each group of 250 been extracted from each of the samples, the 100,000 sample means would be normally distributed, and their mean would be the average test score which would have been attained had the entire population of employable adults taken the test.

*z*-scores

Practical considerations generally preclude examination of a large number of individual samples and thus bits of data usually are extracted from only one sample. However, the information gained from examining one sample can be used to develop information about the entire population from which it was extracted, leading in turn to the probability that a given assertion regarding this population is false.

The examined sample at hand can be construed as an element of the set of all possible samples of like size that could be taken from the given population.\(^7^2\) For example, the first sample taken in

---

\(^7^1\) Unless an entire population is tested, the possibility always exists that the value \(p\) will differ significantly from the true proportion of the population, but the chances of that happening are infinitesimal, about as great as a monkey sitting down by a typewriter and, by randomly striking the keys, typing the text of the United States Constitution. It is not a possibility given any consideration by statisticians.

\(^7^2\) If \(N\) is the size of a population, and \(n\) is the size of a sample to be extracted, then
Illustration 1, which yielded a proportion of 80% passing, could be viewed as an element of the set of the 100,000 such samples taken, which are but a portion of all such samples which might have been extracted from the given population. Similarly, a statistical parameter extracted from the sample at hand can be viewed as one of the many such parameters embodied in each of the theoretically possible samples which are normally distributed. Thus, the 80% proportion taken from the first sample in Illustration 1 can be viewed as an element of the normally distributed set of all such proportions which might have been taken from the theoretically possible samples.

The measurements of the arithmetic data in a normal or near normal distribution with a known mean and standard deviation can be standardized into z-scores or t-scores. After standardizing the measurement of a given bit of data in such a distribution, it is possible to determine where it lies in relation to other bits of data in the distribution, and thus determine the percentage of the distribution’s elements which fall below or lie above it. In particular, it is possible to determine whether or not it lies in the top or

there are

\[ N(N-1)(N-2)(N-3) \ldots (N-n+1) \]
\[ n(n-1)(n-2)(n-3) \ldots 3 \cdot 2 \cdot 1 \]
such samples. In reality, this is an astronomically large number. From a population of just 100, for example, over 75,000,000 samples of size 5 can be extracted.

73. The z-score of any data point \( x \) in a normal distribution is given by the formula:

\[ z = \frac{x - \text{mean}}{\text{standard deviation}} \]

This formulation standardizes the measurements of data within any normal distribution by assigning any data point to a value corresponding to the number of standard deviation units it is removed from the mean. The distribution of z-scores also will be normal with a mean value of zero and a standard deviation with measure one. Thus, a single group of calculations on z-scores will serve to provide results for any normally distributed set of data. See Freund & Williams, supra note 16, at 226-30; Huntsberger, supra note 1, at 136-37.

74. When the samples in question are relatively small (for convenience, this generally is construed to be samples of size 30 or less), the distribution is nearly, but not exactly normal. Freund & Williams, supra note 16, at 266, 514. The size of the discrepancy depends on one less the size of the samples considered \((n - 1)\), called the number of “degrees of freedom” of the distribution. To compensate for this discrepancy, a t-score rather than a z-score is used. The use of a t-score allows for more latitude in the calculations. The t-score for a given bit of data \( x \) is calculated in the same manner as a z-score:

\[ t = \frac{x - \text{mean}}{\text{standard deviation}} \]

See id. at 266-68; Huntsberger, supra note 1, at 240-42. For a discussion of the appropriateness of the use of small sized samples, see note 65 supra.

75. See Freund & Williams, supra note 16, at 226-28, 267.
bottom 5% or 2 1/2% of the distribution. Because the probability that any one bit of randomly selected data possessing a given property corresponds to the percentage of data points in the set with that property, the probability that any data point taken at random from a normal or near normal distribution is in the top or bottom 5% or 2 1/2% of the distribution is 5% or 2 1/2% respectively.

Test for Significance

Operating at a level of significance of 0.05, the validity of a null hypothesis can be tested against an alternate hypothesis by taking a random sample or samples from a population, extracting the appropriate statistical parameters, calculating the relevant t- or z-score, and determining whether it lies above or below the indicated threshold value separating the region of acceptance of the null hypothesis from the regions which prompt its rejection in favor of the alternate.

76. The table in Appendix A lists the t-score threshold values through 29 degrees of freedom, followed by the z-score threshold values separating the shaded 5% of the distribution from the unshaded 95%, pictured above the respective column listings.

Any bit of data with a t-score or a z-score greater than the appropriate threshold value in columns I or IV lies in the shaded area indicating that it is in the top 5% or 2 1/2% of the distribution respectively, while a bit of data with a t- or z-score less than the appropriate threshold value in columns II or III is in the bottom 5% or 2 1/2% of the distribution respectively. For example, given a sample of 16 observations whose mean has a t-score of 1.825, one is able to determine that the sample's mean lies in the top 5% of the distribution because at 15 degrees of freedom (16 - 1), see note 74 supra, the appropriate threshold value taken from column I in Appendix A is 1.753, which is less than 1.825. One also can conclude that this sample's mean is not in the top 2 1/2% of the distribution because 1.825 is not greater than 2.131, the appropriate threshold value taken from column IV.

77. For example, the probability that a male will be selected at random from a group consisting of 80% men and 20% women is 80%.

78. All illustrations and theory presented in this Article shall assume a level of significance of 0.05, but similar tables to the one presented in Appendix A can be constructed for any other level, and the procedures developed herein using the 0.05 level would hold for these others as well.

79. It is the alternate hypothesis which determines the column in Appendix A from which the appropriate threshold value(s) is extracted. If the alternate hypothesis postulates that a sample parameter exceeds that of a population, or that a parameter of a first sample exceeds that of a second, then the appropriate threshold value will be found in column L. If the relevant t- or z-score is greater than this value, then the null hypothesis is to be rejected in favor of the alternate; otherwise it is to stand. Conversely, if the alternate hypothesis postulates that a sample parameter falls below that of a population, or that a parameter of a first sample is exceeded by that of a second, then the appropriate threshold value will be found in column II in Appendix A, and if the t- or z-score falls below this value, the null hypothesis is to be rejected in favor of the alternate. If it exceeds the threshold value, the
Illustration 2: Assume that an employer claims that the average or mean wage paid to white workers is the same as that paid to their black counterparts (a null hypothesis) and that because of the large size of the workforce it would be impractical to attempt to determine the race and wage of each employee. A court faced with the challenge that blacks are paid less on the average will test the null hypothesis that the average wages of the two groups are the same against the alternate hypothesis that on average the white workers are paid more. Because the alternate hypothesis postulates that a first statistical parameter will exceed a second, attention will be focused on column I in Appendix A.80 A random sample of white employees and a random sample of black employees are taken, and the mean wage is calculated for each. The z-score for the difference of these means is then examined. If it is not greater than 1.645, it lies in the unshaded region of acceptance, and the null hypothesis stands, no case of discrimination against the black employees having been established. If it is greater than 1.645, it lies in the shaded region, and the null hypothesis should be rejected in favor of the alternate, thus establishing a prima facie showing of discrimination.

Having established the statistical foundation for hypothesis testing, the remainder of this Article focuses on its use in establishing a prima facie case of employment discrimination and the superiority of this statistical theory to the techniques currently used by courts and the EEOC.

null hypothesis will stand. Finally, if the alternate hypothesis merely postulates a difference, then attention is focused on columns III and IV, and the appropriate threshold values are chosen from them. If the relevant t- or z-score falls within the shaded area (below the threshold value given in column III or above that given in column IV), the null hypothesis is to be rejected and the alternate accepted. If, however, it lies in the unshaded region, then the null hypothesis will stand.

80. See note 79 supra.

81. Assume that each sample contains over 30 workers.

82. The formula used to calculate the z-score for the difference between two means, discussed in the section on Means, see notes 120-29 & accompanying text infra, is:

\[ z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \]

where \( \bar{x}_1, s_1 \) and \( n_1 \) are the mean, standard deviation and size of the first sample respectively, and \( \bar{x}_2, s_2 \) and \( n_2 \) are those respective parameters of the second sample.

83. The number 1.645 is the threshold value for the z-score in column I.
Proportions

Employers can discriminate against specific minorities by developing a workforce where the proportion of minority members is significantly smaller than the proportion of minority members in the population of available employees, or by establishing criteria for employment or promotion which tend to exclude a greater proportion of minority workers than those of other groups within the population. Requiring a complete mathematical mirroring of the relevant proportions is not feasible given the circumstances surrounding the construction of a workforce. Rather, the issue is whether the disparity between these proportions is so great as to warrant an inference of discrimination. Formulae for calculating t- and z-scores will be introduced and tailored to each specific situation, enabling a court, through the use of hypothesis testing, to resolve the issue and determine whether a prima facie case has been established.

Test for Proportions

To determine whether a specific minority is significantly underrepresented in a defendant's workforce, a null hypothesis—that no significant difference exists between the proportion of minority employees in the workforce and the proportion of minority workers in the population of available employees—is tested against the alternate hypothesis that the proportion of minority employees in the workforce is significantly less than that of the population. The t- or z-score formula for this test is:

\[ z = \frac{x - p}{\sqrt{\frac{p(1 - p)}{n}}} \]

or equivalently:

\[ z = \frac{x - np}{\sqrt{np(1 - p)}} \]

84. See note 19 supra.
85. See note 20 supra.
86. See Freund & Williams, supra note 16, at 324.
where \( x \) is the number of minority members, \( n \) is the total number of employees in the defendant’s workforce, and \( p \) is the proportion of the population that are minority members. If this score falls below the appropriate threshold value found in column II of Appendix A, the null hypothesis is to be rejected in favor of the alternative hypothesis. If, however, the t- or z-score lies above the appropriate threshold value, the null hypothesis will stand.

**Illustration 3:** Assume that of a workforce of 5,000 \((n)\), 36% or 1,800 employees are black \((x)\), while the proportion of blacks in the population of available employees is 40%. Then:

\[
z = \frac{1,800 - (5,000)(0.40)}{\sqrt{(5,000)(0.40)(1 - 0.40)}} = \frac{1,800 - 2,000}{\sqrt{1,200}} = \frac{-200}{34.64} = -5.77.
\]

Because this value is less than \(-1.645\), the null hypothesis postulating no significant difference between the relevant proportions must be rejected in favor of the alternate, and a prima facie case of discrimination is established.

It should be noted from the results in Illustration 3 that the court would be at least 95% certain that blacks were significantly underrepresented in the defendant’s workforce even though the disparity between the proportion of blacks in the defendant’s employ (0.36) and the proportion of blacks in the population of available employees (0.40) was only 4%. Comparing percentage levels alone is not sufficient to determine whether such a disparity is significant except in the most extreme circumstances. Even then,

---

87. Because the alternate hypothesis postulates that the sample proportion is less than the population proportion, the appropriate threshold value is found in column II. See note 79 supra.

88. Because \( n \) is greater than 30, the appropriate threshold value is the z-value \(-1.645\). See Appendix A infra.

89. Consider an employer with a workforce of 200 \((n)\) of which 72 \((x)\), or 36% are black. Let the proportion of blacks in the population of available employees be 0.40. Thus, the proportion of blacks in the workforce and in the population in this example match those in Illustration 3. Then:

\[
z = \frac{72 - (200)(0.40)}{\sqrt{(200)(0.40)(1 - 0.40)}} = \frac{72 - 80}{\sqrt{48}} = \frac{-8}{6.93} = -1.15.
\]

This score is greater than the threshold value \(-1.645\). Therefore, the null hypothesis stands, and, contrary to the conclusion reached in Illustration 3, no case that blacks are significantly underrepresented in this workforce has been established.

90. In New Orleans Pub. Serv., Inc. v. Brown, 507 F.2d 160, 163-64 (5th Cir. 1975), the charge filed by the EEOC alleged that the defendant employed 695 blacks \((x)\) out of a total workforce of 3,052 \((n)\), and the relevant population \((p)\) was 45% black. It was clear to the court that the disparity between the proportion of blacks in the workforce, 695/3,052 =
the theory of hypothesis testing will reach the same result.91

**Difference in Proportions Test**

Employers may also discriminate against a specific minority by establishing criteria for employment or promotion which exclude or disqualify a greater proportion of that minority than those of other groups.92 When complete information is available with respect to the percentages of those in the relevant groups who were excluded, then the conclusion of whether such discrimination exists is reached merely by comparing these percentages and observing whether the specific minority members were excluded in greater proportions.93 Absent such complete information, the information at hand must be used in making this determination. If only sample data are available, the disparity between the sample proportions must be examined and a determination must be made as to whether it is large enough to warrant an inference of discrimination.

The null hypothesis that there is no significant difference in the proportions of the relevant minority and nonminority groups disqualified or excluded by the challenged criteria is tested against the alternate hypothesis that the relevant minority members are excluded in greater proportions. The z-score94 for this test is:

\[
z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}},
\]

where \(x_1\) is the number of minority applicants rejected, \(n_1\) the total number of minority applicants, \(x_2\) the number of nonminority applicants rejected, and \(n_2\) is the total number of nonminority appli-

22.8%, and the proportion of blacks in the population was significant by mere inspection of figures.

91. In New Orleans Pub. Serv., Inc. v. Brown, 507 F.2d 160 (5th Cir. 1975), see note 90 supra, the relevant \(x\), \(n\), and \(p\) values yielded a z-score of \(-24.69\), which is less than the threshold value \(-1.645\). This implies that the null hypothesis fails in favor of the alternate hypothesis that the proportion of blacks in the New Orleans Public Service workforce is significantly less than the proportion of blacks in the population.

92. See note 20 supra.


94. Without further restrictions on the characteristics of the population, each sample must contain at least 30 observations.

95. See *Freund & Williams*, supra note 16, at 332, problem 5; *Huntsberger*, supra note 1, at 302.
cants. The proportion of applicants who would have been rejected assuming that the null hypothesis is true is represented by \( p \), making \((1 - p)\) the representative of the proportion of the applicants who would have passed. The proportion represented by \( p \) equals the total number of applicants rejected divided by the total number of applicants:

\[
p = \frac{x_1 + x_2}{x_1 + x_2}.
\]

If this z-score is greater than 1.645,\(^{97}\) the null hypothesis is rejected; if less than 1.645, no case of discriminatory treatment has been established.

**Illustration 4:** Assume that a written aptitude test is administered to 800 women (\( n_1 \)) and 1,200 men (\( n_2 \)) as part of a company's hiring procedure. Assume further that 400 women (\( x_1 \)) and 500 men (\( x_2 \)) fail the test, and a Title VII action is brought against the company alleging that the test had a discriminatory impact on women. Testing the null hypothesis that the test was not discriminatory against the alternate hypothesis that the proportion of women failing is significantly greater than the proportion of men failing,

\[
p = \frac{400 + 500}{800 + 1,200} = \frac{900}{2,000} = 0.45.
\]

Thus, \( 1 - p = 0.55 \), and

\[
z = \frac{400}{800} - \frac{500}{1,200} = \frac{0.500 - 0.417}{\sqrt{\frac{0.45(0.55)}{800} + \frac{(0.45)(0.55)}{1,200}}} = \frac{0.083}{0.023} = 3.61.
\]

Because this score is greater than 1.645, the null hypothesis is rejected, and a prima facie case of discrimination is established.

Unlike a simple comparison of proportions which might suffice in extreme cases,\(^{98}\) this test can be relied on in all instances.\(^{98}\) The

\(^{96}\) Id.

\(^{97}\) Because the alternate hypothesis postulates that the first sample proportion is greater than the second sample proportion, the appropriate threshold value is found in column I of Appendix A. See note 79 supra.

\(^{98}\) In United States v. Commonwealth of Va., 454 F. Supp. 1077 (E.D. Va. 1978), 145 blacks took a written aptitude test for Virginia State trooper positions in 1973-1975 (\( n_1 \)), and 101 failed (\( x_1 \)). During this period, 655 whites took the same test (\( n_2 \)), and 160 failed (\( x_2 \)). The court saw by mere inspection of the data that nearly 70% of the blacks failed while less than 25% of the whites failed. This disparity was of great enough significance for the court to pass the burden of proving the test's validity to the defendants.
difference in proportions test also can be used to determine whether the proportion of a given minority in a defendant's workforce in lower paying and less desirable jobs is significantly greater than the proportion of nonminority employees in comparable positions.100

The EEOC Four-Fifths Rule

Instead of the difference in proportions test, the EEOC and some courts have adopted a standard known as the "four-fifths rule"101 in which a difference between two proportions is not considered significant if the proportion of the successes of the relevant minority sample is at least four-fifths, or 80%, of the proportion of the successes of the nonminority sample. Thus, if \( r_m \) is the proportion of the relevant minority sample that satisfied the challenged employment or promotion criteria or are in better paying or more desirable jobs, and \( r_w \) is the nonminority sample proportion in that

99. Assume that an aptitude test is administered to 80 women and 120 men, and 40 of the women and 50 of the men fail. The relevant proportions of the women and men who failed this test match those in Illustration 4, as does the null hypothesis representative of the proportion who would have failed. Yet in this example:

\[
Z = \frac{(0.45)(0.55) + (0.45)(0.55)}{\sqrt{\frac{80}{0.005156}} + \frac{120}{0.083}} = 1.15.
\]

Because the \( z \)-score is less than 1.645, the null hypothesis must stand as there has been no showing of a significant disparity between the proportions.

100. In Abron v. Black & Decker Mfg. Co., 439 F. Supp. 1095, 1102 (D. Md. 1977), the court found, inter alia, that 52 of the 117 blacks employed by the defendant were assigned to the less desirable departments, while 593 of the 2,926 whites employed were so assigned. Using these facts:

\[
Z = \frac{52 + 593}{117 + 2,926} = \frac{645}{3,043} = 0.212,
\]

\[
1 - p = 1 - 0.212 = 0.788, \text{ and}
\]

\[
Z = \frac{52}{117} - \frac{593}{2,926} = \frac{0.24}{0.04} = 6.0.
\]

Because this score was greater than 1.645, a prima facie case of discrimination was established.

position, then the EEOC would rule that no case of discrimination had been advanced if \( \frac{r_m}{r_w} \) is greater than 80%. Conversely, the EEOC would find a prima facie case of discrimination if the ratio is less than 80%.

This rule can easily lead to inaccurate results. In Illustration 4, for example, 400 out of 800 women and 700 out of 1,200 men who took the challenged test passed. Therefore, \( r_m = 400/800 = 0.500 \), and \( r_w = 700/1,200 = 0.583 \). Because the ratio \( \frac{r_m}{r_w} = 0.500/0.583 = 0.86 \) (greater than 80%) the EEOC would conclude, contrary to the weight of the evidence under hypothesis testing, that the employment test described did not have a disparate impact on women. In some cases the weight of the evidence may not establish a prima facie case of discrimination under hypothesis testing, but the ratio \( \frac{r_m}{r_w} \) will nonetheless be less than 80%. The difference in proportions test clearly is superior to the less precise four-fifths rule, and the latter should be discarded in favor of the more sophisticated mathematical technique.

The \( \chi^2 \) Test

The difference in proportions test, albeit sound, is nonetheless limited in that it cannot be used in cases considering more than two proportions. When there are more than two relevant groups in the population or more than two relevant categories of placement of employees or potential employees, the difference in proportions test is inapplicable, and the \( \chi^2 \) (Chi Square) Test should be employed.

102. This limitation of the four-fifths rule is analyzed thoroughly by Professor Shoben. Comment, Differential Pass-Fail Rates in Employment Testing: Statistical Proof Under Title VII, 91 Harv. L. Rev. 793, 805-12 (1978). See also id. at 809 n.59 (Employer B).

103. Id. at 809 n.56 (Employer D).

104. At times, consideration may have to be given to more than two distinct groups within the population. See, e.g., Western Addition Community Organization v. Alioto, 340 F. Supp. 1351 (N.D. Cal. 1972) (court considered three groups: whites, blacks and Mexican-Americans).

105. The categories considered in the difference in proportions test subsection have been in pairs such as pass-fail or hired-excluded. See notes 92-100 and accompanying text supra. Yet at times, consideration may have to be given to various levels or types of jobs, and more than two categories would have to be investigated. See, e.g., Garrett v. R.J. Reynolds Indus., Inc., 81 F.R.D. 25, 34-35 (M.D.N.C. 1978) (job categories examined included officials and managers, professionals, technicians, office and clerical, craftsmen (skilled), and operators (semi-skilled)).

The $X^2$ Test is an extension of the techniques of testing proportions previously discussed. As with the previous tests discussed, a null hypothesis—no significant discrepancy exists in the treatment of the various groups considered—is tested against an alternate hypothesis—that a significant disparity exists to warrant an inference of such a discrepancy—by calculating a number or score (in this case $X^2$) and determining whether it lies above or below a specific threshold value. To begin the calculation, representative values for the relevant categorical proportions are determined. This is accomplished by taking the total number of employees or applicants in each category considered, and dividing by the total number of employees or applicants. These representative proportions are then multiplied by the number of employees or applicants in each group to yield the respective expected values. These are the expected numbers of members from each relevant group who would populate each of the considered categories if the null hypothesis was true. They are compared to the actual or observed numbers of members from the respective groups populating the respective categories and, if the disparity between these values is great enough, the null hypothesis will be rejected in favor of the alternate hypothesis.

The $X^2$ disparity is computed by taking the squares of the differences between the corresponding expected and observed values in each category with respect to each group, dividing by the respective expected value, and adding all of these scores. If this number is greater than the appropriate $X^2$ threshold value found in Appendix A, the null hypothesis is rejected, and a prima facie case of disparate treatment has been advanced. If the number is less than the threshold value, the null hypothesis will stand.

The relevant threshold value is determined by multiplying one less the number of groups considered by one less the number of

107. In fact, the $X^2$ Test also may be used in lieu of the difference in proportions test when there are only two groups and two categories being considered. See Freund & Williams, supra note 16, at 333 (problem 6). However, the preference is to use the difference in proportions test whenever possible because it is considered a more simple and direct procedure to employ, understand, and explain. It is important to note, however, that the difference in proportions test is not a more mathmatically sound theory than the $X^2$ Test.

108. See Freund & Williams, supra note 16, at 328-29.

109. Id.

110. Id. at 330.

111. Id. at 331.

112. Id.
categories to attain the appropriate number of degrees of freedom\textsuperscript{113} and then extracting the corresponding value from column V in Appendix A.\textsuperscript{114}

Illustration 5: Assume that a defendant company administered a preemployment test to 5,200 whites, 3,200 blacks, and 1,600 Spanish-surnamed individuals and 3,000 whites, 2,000 blacks, and 1,000 Spanish-surnamed individuals failed the test. Assume further that the test was challenged as having a disparate impact on minorities. From this information, three groups (white, black, Spanish-surnamed) and two categories (pass, fail) are considered. The following table can be constructed:

<table>
<thead>
<tr>
<th>Groups:</th>
<th>whites</th>
<th>blacks</th>
<th>Spanish-surnamed</th>
<th>Total in Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categories:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passed</td>
<td>2,200</td>
<td>1,200</td>
<td>600</td>
<td>4,000</td>
</tr>
<tr>
<td>Failed</td>
<td>3,000</td>
<td>2,000</td>
<td>1,000</td>
<td>6,000</td>
</tr>
<tr>
<td>Total in Group:</td>
<td>5,200</td>
<td>3,200</td>
<td>1,600</td>
<td>10,000</td>
</tr>
</tbody>
</table>

Let \( p \) be the representative proportion of those who would pass and \( q \) of those who would fail (the relevant categorical proportions), assuming the null hypothesis to be true.

\[
p = \frac{\text{total number of applicants who passed}}{\text{total number of applicants}} = \frac{4,000}{10,000} = 0.40. \\
q = \frac{\text{total number of applicants who failed}}{\text{total number of applicants}} = \frac{6,000}{10,000} = 0.60.
\]

The number of whites, blacks and Spanish-surnamed individuals expected to pass and fail the test (the expected values) are determined by multiplying these representative proportions by the number of applicants in each group:

- The number of whites expected to pass = \((0.40) (5,200) = 2,080.\)
- The number of whites expected to fail = \((0.60) (5,200) = 3,120.\)
- The number of blacks expected to pass = \((0.40) (3,200) = 1,280.\)
- The number of blacks expected to fail = \((0.60) (3,200) = 1,920.\)

\textsuperscript{113} Id. at 330.
\textsuperscript{114} These threshold values, like the others in Appendix A, correspond to a level of significance of 0.05.
The number of Spanish-surnamed individuals expected to pass = 
(0.40) (1,600) = 640.

The number of Spanish-surnamed individuals expected to fail = 
(0.60) (1,600) = 960.

These expected values are tested against the corresponding observed values in the sample. Thus

\[ X^2 = \text{the sum of } \frac{(\text{expected values} - \text{observed values})^2}{\text{expected values}} \]

\[ = \frac{(2,080 - 2,200)^2}{2,080} + \frac{(3,120 - 3,000)^2}{3,120} + \frac{(1,280 - 1,200)^2}{1,280} + \frac{(1,920 - 2,000)^2}{1,920} + \frac{(640 - 600)^2}{640} + \frac{(960 - 1,000)^2}{960} \]

\[ = \frac{14,400}{2,080} + \frac{14,400}{3,120} + \frac{6,400}{1,280} + \frac{6,400}{1,920} + \frac{1,600}{640} + \frac{1,600}{960} \]

\[ = 6.92 + 4.62 + 5.00 + 3.33 + 2.50 + 1.67 = 24.04. \]

Because there are three groups and two categories, one less the number of groups multiplied by one less the number of categories, \((3 - 1) \times (2 - 1) = 2 \times 1\), yields two degrees of freedom. Thus, the relevant threshold value is 5.991.\(^{116}\)

Here, the \(X^2\) value of 24.04 is greater than 5.991. The null hypothesis that there is no significant disparity in the proportions therefore is rejected in favor of the alternate hypothesis that the test does have a disparate impact.\(^{117}\) By observation one can see that whites did better than expected while blacks and Spanish-surnamed individuals did worse than anticipated. Therefore a prima facie case of discrimination against nonwhites has been established.

As before, inspection of the data and percentage comparisons alone may lead to inaccurate results.\(^{118}\) Additionally, combining
groups or categories so as to produce two of each, and then using the difference in proportions test, may cause inaccurate results. The $X^2$ Test is a sound method for determining whether a disparity in treatment of specific groups is large enough to warrant an inference of discrimination and should always be used if there are more than two groups or two categories to be considered.

**Means**

Examination and comparison of proportions is not the only way hypothesis testing can be used in establishing or rebutting a prima facie case in a Title VII action. Courts also have compared relevant minority and nonminority sample means in order to determine whether discrimination in employment exists. The focus

matching the representative proportions in Illustration 5. Thus, the expected values are:

- whites passing: $(0.40)(520) = 208$.
- whites failing: $(0.60)(520) = 312$.
- blacks passing: $(0.40)(320) = 128$.
- blacks failing: $(0.60)(320) = 192$.
- Spanish-surnamed individuals passing: $(0.40)(160) = 64$.
- Spanish-surnamed individuals failing: $(0.60)(160) = 96$.

Thus:

$$X^2 = \frac{(208 - 220)^2}{208} + \frac{(312 - 300)^2}{312} + \frac{(128 - 120)^2}{128} + \frac{(192 - 200)^2}{192} + \frac{(64 - 60)^2}{64} + \frac{(96 - 100)^2}{96}$$

$$= 0.692 + 0.462 + 0.500 + 0.333 + 0.250 + 0.167$$

$$= 2.404.$$  
As in Illustration 5, there are two degrees of freedom. Thus, the threshold value is again 5.991. However, contrary to the results in Illustration 5, $X^2$ is less than the threshold value, implying that the null hypothesis will stand, and no case of discrimination has been established.

119. Assume that a preemployment test is administered to 100 whites, 100 blacks, and 100 Mexican-Americans, and that 50 whites, 35 blacks, and 50 Mexican-Americans pass, while the remaining applicants failed. Computing $X^*$, one gets a value of 6.061. Comparing it to the relevant threshold value, 5.991, one can conclude that a significant disparity exists. Inspection of the data shows that it is blacks who were disadvantaged.

However, if blacks and Mexican-Americans were combined to form one group of minority applicants, the difference in proportions test can be used in comparing the proportion of white failures (50/100 or 0.50) with the proportion of minority failures (115/200 or 0.575). The z-score obtained is 1.23 which is less than the relevant threshold value 1.645. Thus, one would be unable to reject the null hypothesis and conclude that blacks were being disadvantaged.

120. See notes 121-24 infra.
is usually on the differences between sample means of salaries,\(^{121}\) increments,\(^{122}\) seniority or tenure before promotions,\(^{123}\) and test scores and merit ratings\(^{124}\) among the workforce within the defendant company in order to determine whether a particular minority was being discriminated against.

Use of hypothesis testing in these instances entails considering a null hypothesis that there is no significant difference between the means of the relevant minority and nonminority population groups, which is tested against an alternate hypothesis postulating that a significant difference exists disadvantaging the minority employees. Such disadvantage with respect to salaries, increments and merit ratings is that the minority mean is significantly less than the corresponding nonminority mean, and with respect to tenure before promotion, that it is significantly greater than its nonminority counterpart.

Standard deviations, as well as means, must be extracted from the relevant samples, since the \(z\)-score is:

\[
z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}}^{125}
\]

where \(\bar{x}_1\) and \(\bar{x}_2\) are the means of the respective minority and nonminority samples extracted from the workforce, \(s_1\) and \(s_2\) are the respective standard deviations\(^{126}\) of these samples, and \(n_1\) and \(n_2\) are the respective sizes of each sample. The threshold values are found in the same manner as previously discussed in the difference

\(^{121}\) See, e.g., James v. Stockham Valves & Fittings, 559 F.2d 310, 327 (5th Cir. 1977), cert. denied, 434 U.S. 1034 (1978); Watkins v. Scott Paper Co., 530 F.2d 1159, 1165 (5th Cir. 1976).


\(^{125}\) Each sample must contain at least 30 observations, because without imposing further restrictions on the behavior of the relevant statistical parameters, only \(z\)-scores can be used and small sized samples are not considered. See FREUND & WILLIAMS, supra note 16, at 292-94.

\(^{126}\) See note 68 supra.
Illustration 6: Assume that an employer has a workforce consisting of 100 female and 200 male workers in comparable jobs and that the 100 females received pay raises in an increment which averaged $850 based upon various objective and subjective criteria. The standard deviation of these 100 raises was $270. The male workers, with their increments purportedly based on the same criteria, received an average raise of $940, with a standard deviation of $260.

In testing the null hypothesis that this $90 difference in the increments is not significant enough to warrant an inference of discrimination against females against the alternate hypothesis that the difference is significant enough to warrant that inference, the relevant z-score is computed to be:

\[ z = \frac{850 - 940}{\sqrt{\frac{(270)^2}{100} + \frac{(260)^2}{200}}} = \frac{-90}{\sqrt{729 + 338}} = \frac{-90}{32.7} = -2.75. \]

Because -2.75 is less than -1.645, the null hypothesis is rejected in favor of the alternate, and a prima facie case of discrimination has been established.  

Conclusion

Statistical evidence, like any other kind of evidence, is rebuttable. While its value ultimately depends on the circumstances surrounding each case, the techniques developed in this Article, when properly applied, are irrefutable in what they demonstrate. When led to a rejection of the null hypothesis at a level of significance of 0.05, a court can be at least 95% confident that a disparity of treatment of the relevant groups exists. It is, of course, for the courts to determine whether the application of this analysis and theory is correct and proper, but if they so determine, then the

127. See notes 92-100 & accompanying text supra. See also notes 76, 79 supra.
128. The threshold value of -1.645 is found in column II, because the alternate hypothesis postulates that women received a significantly smaller increment than men, and therefore the first value will be less than the second. See note 79 supra.
129. Comparisons between more than two sample means also can be made to determine whether significant differences exist which disadvantage specific minorities, but a discussion of the methods involved is beyond the limited scope of this Article. See FREUND & WILLIAMS, supra note 16, at 344-60; HUNTSBERGER, supra note 1, at 320-44.
130. See note 13 & accompanying text supra.
evidence flowing from the statistics should not be discounted or dismissed. Rather, in light of the increasingly sophisticated treatment of Title VII actions, statistical analysis must be viewed as both a proper and necessary tool for the establishment or refutation of the prima facie case.
## APPENDIX A

<table>
<thead>
<tr>
<th>Degrees of Freedom</th>
<th>(t_{0.05})</th>
<th>(-t_{0.05})</th>
<th>(-t_{0.025})</th>
<th>(t_{0.025})</th>
<th>(X^2_{0.05})</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.920</td>
<td>-2.920</td>
<td>-4.303</td>
<td>4.303</td>
<td>5.991</td>
</tr>
<tr>
<td>3</td>
<td>2.353</td>
<td>-2.353</td>
<td>-3.812</td>
<td>3.812</td>
<td>7.815</td>
</tr>
<tr>
<td>4</td>
<td>2.132</td>
<td>-2.132</td>
<td>-2.776</td>
<td>2.776</td>
<td>9.488</td>
</tr>
<tr>
<td>5</td>
<td>2.015</td>
<td>-2.015</td>
<td>-2.571</td>
<td>2.571</td>
<td>11.070</td>
</tr>
<tr>
<td>6</td>
<td>1.943</td>
<td>-1.943</td>
<td>-2.447</td>
<td>2.447</td>
<td>12.592</td>
</tr>
<tr>
<td>7</td>
<td>1.895</td>
<td>-1.895</td>
<td>-2.365</td>
<td>2.365</td>
<td>14.067</td>
</tr>
<tr>
<td>8</td>
<td>1.860</td>
<td>-1.860</td>
<td>-2.306</td>
<td>2.306</td>
<td>15.507</td>
</tr>
<tr>
<td>9</td>
<td>1.833</td>
<td>-1.833</td>
<td>-2.262</td>
<td>2.262</td>
<td>16.919</td>
</tr>
<tr>
<td>10</td>
<td>1.812</td>
<td>-1.812</td>
<td>-2.228</td>
<td>2.228</td>
<td>18.307</td>
</tr>
<tr>
<td>11</td>
<td>1.793</td>
<td>-1.793</td>
<td>-2.201</td>
<td>2.201</td>
<td>19.675</td>
</tr>
<tr>
<td>12</td>
<td>1.782</td>
<td>-1.782</td>
<td>-2.179</td>
<td>2.179</td>
<td>21.026</td>
</tr>
<tr>
<td>13</td>
<td>1.771</td>
<td>-1.771</td>
<td>-2.160</td>
<td>2.160</td>
<td>22.362</td>
</tr>
<tr>
<td>14</td>
<td>1.761</td>
<td>-1.761</td>
<td>-2.145</td>
<td>2.145</td>
<td>23.685</td>
</tr>
<tr>
<td>15</td>
<td>1.753</td>
<td>-1.753</td>
<td>-2.131</td>
<td>2.131</td>
<td>24.996</td>
</tr>
<tr>
<td>16</td>
<td>1.746</td>
<td>-1.746</td>
<td>-2.120</td>
<td>2.120</td>
<td>26.296</td>
</tr>
<tr>
<td>17</td>
<td>1.740</td>
<td>-1.740</td>
<td>-2.110</td>
<td>2.110</td>
<td>27.587</td>
</tr>
<tr>
<td>18</td>
<td>1.734</td>
<td>-1.734</td>
<td>-2.101</td>
<td>2.101</td>
<td>28.869</td>
</tr>
<tr>
<td>19</td>
<td>1.729</td>
<td>-1.729</td>
<td>-2.093</td>
<td>2.093</td>
<td>30.144</td>
</tr>
<tr>
<td>20</td>
<td>1.725</td>
<td>-1.725</td>
<td>-2.086</td>
<td>2.086</td>
<td>31.410</td>
</tr>
<tr>
<td>21</td>
<td>1.721</td>
<td>-1.721</td>
<td>-2.080</td>
<td>2.080</td>
<td>32.671</td>
</tr>
<tr>
<td>22</td>
<td>1.717</td>
<td>-1.717</td>
<td>-2.074</td>
<td>2.074</td>
<td>33.924</td>
</tr>
<tr>
<td>23</td>
<td>1.714</td>
<td>-1.714</td>
<td>-2.069</td>
<td>2.069</td>
<td>35.172</td>
</tr>
<tr>
<td>24</td>
<td>1.711</td>
<td>-1.711</td>
<td>-2.064</td>
<td>2.064</td>
<td>36.415</td>
</tr>
<tr>
<td>25</td>
<td>1.708</td>
<td>-1.708</td>
<td>-2.060</td>
<td>2.060</td>
<td>37.652</td>
</tr>
<tr>
<td>26</td>
<td>1.706</td>
<td>-1.706</td>
<td>-2.056</td>
<td>2.056</td>
<td>38.885</td>
</tr>
<tr>
<td>27</td>
<td>1.703</td>
<td>-1.703</td>
<td>-2.052</td>
<td>2.052</td>
<td>40.113</td>
</tr>
<tr>
<td>28</td>
<td>1.701</td>
<td>-1.701</td>
<td>-2.048</td>
<td>2.048</td>
<td>41.337</td>
</tr>
<tr>
<td>29</td>
<td>1.699</td>
<td>-1.699</td>
<td>-2.045</td>
<td>2.045</td>
<td>42.557</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Degrees of Freedom</th>
<th>(t_{0.025})</th>
<th>(-t_{0.05})</th>
<th>(-t_{0.025})</th>
<th>(t_{0.05})</th>
<th>[z_{0.05}]</th>
</tr>
</thead>
<tbody>
<tr>
<td>inf.</td>
<td>1.645</td>
<td>-1.645</td>
<td>-1.960</td>
<td>1.960</td>
<td>inf.</td>
</tr>
</tbody>
</table>

\(t_{0.05}\), \(-t_{0.05}\), \(-t_{0.025}\), \(t_{0.025}\), \[z_{0.05}\], \[z_{0.025}\], \[X^2_{0.05}\]