“Gross Statistical Disparities” as Evidence of a Pattern and Practice of Discrimination: Statistical versus Legal Significance

Allan G. King*

I. Introduction

The “pattern and practice” theory of discrimination is a linchpin of employment discrimination class actions. Because this theory alleges that discrimination is an employer’s “standard operating procedure”—its usual, rather than its exceptional, way of doing business—it weaves together the many individual decisions affecting an employer’s workforce into a pattern that a court can adjudicate in a single class-action proceeding.1

Pattern and practice cases typically proceed in two stages. In Stage I, the fact-finder answers whether the plaintiff has proven a pattern and practice of discrimination by a preponderance of the evidence.2 If the plaintiff prevails in Stage I, then in Stage II the court applies a presumption that the employment decision affecting each class member resulted from the discriminatory practice.3 The burden thus shifts to the employer to prove that, notwithstanding its presumptively discriminatory pattern and practice, its decision regarding any particular class member was nondiscriminatory.4

The Supreme Court has explained that proof of a pattern and practice primarily depends on statistical evidence, perhaps supplemented with anecdotal testimony that brings the statistical evidence “convincingly to life.”5 Because the pattern and practice finding raises a presumption that the defendant employer subjected each member of the class to discrimination, the Court requires plaintiffs to present

---

*Mr. King is a shareholder in the Dallas office of Littler Mendelson, P.C. He co-chairs the firm’s Class Action Practice Group. Jeremy W. Hawpe, an associate in the firm’s Dallas office, provided helpful research assistance.

2. Id. at 360.
3. Id.
4. Id. at 360, 362.
5. Id. at 339.
In the leading case of Teamsters v. United States, the Court observed that the “inexorable zero” evidenced the total exclusion of African Americans from over-the-road trucking jobs, and thereby was probative of plaintiff’s allegations. Although the Court continued to insist that statistical proof of a pattern and practice required evidence of gross disparities, the statistical evidence in Teamsters and its subsequent pattern and practice cases were so lopsided that there was no need for the Court to identify the statistical threshold that corresponded to this “gross” requirement.

In the absence of a definitive ruling, lower courts frequently have turned to “statistical significance” as the measuring rod. This paper contends that uncritical reliance on statistical significance as a measure of “grossness” is misguided, because, other things being equal, statistical significance depends on the size of the sample, or, in the context of discrimination law, the size of the employer’s labor force. As a result, statistical significance conflates the magnitude of any disparity (e.g., the “shortfall” in the number of protected group members promoted) with the number of employment decisions at issue. In other words, if applied mechanistically, “statistical significance” can mislead by indicating disparities at Employer A are “grosser” than those at Employer B, when in fact the opposite is true, merely because Employer A is much larger.

Additionally, the “bright line” rules established by lower courts have been applied inconsistently. Courts deciding single-plaintiff cases routinely insist that generalized statistical proof, even if statistically significant, lacks probative value if the employer meets its burden of production and advances an individualized explanation for its conduct. Yet, in a pattern and practice case, the same quantum of statistical proof may satisfy the employee’s burden of persuasion, not merely with respect to the plaintiff, but to each and every class member as well.

In light of these shortcomings, this paper contends that courts deciding pattern and practice discrimination cases have been far too

---

accepting of statistical evidence that, in other contexts, they might view as merely marginal. Instead, courts should consider, for example, the statistical principles honed in toxic tort cases (particularly those in which epidemiological studies play a primary role). This comparison is apt because at issue in pattern and practice cases is whether a disparity between any two demographic groups, in terms of hiring, promotion, termination, etc., more likely results from actionable discrimination than from benign, nondiscriminatory causes.

Statistical evidence in pattern and practice cases poses the same conundrum that arises in toxic tort cases: because the phenomenon under study—in tort cases, perhaps a disease, in discrimination cases, perhaps a termination—occurs with some frequency even among those with no exposure to the causative agent (toxicity or discrimination), how much more often must it occur in the exposed group to prove that any injury more likely than not was wrongfully caused? In the parlance of the Supreme Court, how “gross” must statistical disparities be to prove a pattern and practice of discrimination?12

II. The Pattern and Practice Theory

A pattern and practice case alleges that systemic discrimination pervades the workplace—in other words, that it is the employer's standard business practice. If proved, the court deems that a presumptively discriminatory motive taints each decision involving a member of the protected group. The burden then shifts to the employer to establish that, notwithstanding this apparent systemic discrimination, it would have made the same decision with respect to any given employee.13

This theory is well-suited to class actions because it identifies a common practice that is presumed to affect each class member, thereby demonstrating “commonality”; and if the named plaintiffs can convincingly attribute their injuries to the discriminatory practice, then a court readily may find that these plaintiffs’ situations are “typical” of the harm associated with the challenged practice.14 Indeed, many circuits hold that a pattern and practice claim is viable only when pled on behalf of a class.15 Thus, it is not surprising that pattern and practice

---

14. The four requirements for class certification that a plaintiff must satisfy under Rule 23(a) are numerosity, commonality, typicality, and adequacy of representation. Fed. R. Civ. P. 23(a).
claims are a mainstay of class action litigation, or that these claims often produce headline-grabbing settlements.

Historically, plaintiffs have most easily established such patterns and practices when the employer has maintained an express policy of discrimination. For example, in the early Title VII cases, litigation frequently targeted facially discriminatory policies, such as segregated seniority rosters or blanket racial exclusions from particular positions.\(^\text{16}\) In such cases, statistical evidence merely confirms that the policies at issue had their intended discriminatory effect, i.e., the “inexorable zero” is the proof that the challenged policy succeeded in excluding African Americans.

In more recent years, plaintiffs have challenged policies that are, at best, implicitly discriminatory. Thus, plaintiffs may allege that the employer permits a culture of discrimination to flourish, and that this culture has tinged with discrimination all resulting employment decisions.\(^\text{17}\) Frequently, class plaintiffs have coupled such claims with allegations of excessively subjective corporate decision making, which permits free reign to discriminatory influences.\(^\text{18}\)

In such cases, the parties will hotly dispute whether unbridled decision making and a discriminatory corporate culture exist at all, as well as any connection between the culture and the pattern of employment decisions alleged. Statistics, therefore, play an even more central role in challenging these implicit policies and practices than in lawsuits attacking explicit policies. In cases attacking implicit policies, plaintiffs may offer statistical proof of widespread disparities to demonstrate both that the implied discriminatory policy exists and that the resulting disparities are sufficiently large to evidence a widespread discriminatory intent, i.e., a pattern and practice.

### III. The Prevailing Standard of Statistical Proof in Pattern and Practice Cases

Given the primacy of statistical evidence in establishing a pattern and practice of discrimination, and the resulting shift in the burden of persuasion once a plaintiff has made this proof, the standard courts apply in determining the sufficiency of this evidence is critically important. Citations to the line of cases that supply the criterion employed by most courts generally begin with the Supreme Court’s decision in *Castaneda v. Partida*.\(^\text{19}\) Yet *Castaneda* was not an employment

---


\(^{19}\) 430 U.S. 482 (1977).
discrimination case, much less a pattern and practice case. At issue was whether a South Texas county’s method of convening a grand jury unfairly excluded Mexican Americans, resulting in discriminatory adjudication of Mexican-American defendants in criminal cases.20

In reviewing this claim, the Court compared the percentage of Mexican Americans, among those summoned to serve on the county’s grand juries, to Mexican-American representation among the county’s eligible population.21 The Court noted that there were only 339 Mexican Americans among 870 grand jurors summoned during the relevant timeframe, and that strictly proportional representation would have seated 688 Mexican-American grand jurors.22 The Court considered this disparity of nearly 100% to be material, observing that social scientists generally concern themselves when disparities exceed two or three standard deviations, yet the difference between the actual and expected numbers of Mexican-American grand jurors exceeded twenty-nine standard deviations.23 Based in part on that comparison, the Court affirmed the district court’s finding that Mexican Americans were discriminatorily excluded from grand jury service.24

The Court again referenced the benchmark of “two or three standard deviations” in Hazelwood School District v. United States.25 Hazelwood was a pattern and practice suit alleging that a school district engaged in the discriminatory hiring of African American teachers.26 The Court compared the percentage of teachers in the district who were African Americans to the percentage in the relevant labor market.27 Noting that the disparity exceeded six standard deviations in one year and five standard deviations in the following year, the Court concluded that the statistical evidence reflected a “gross” disparity that was probative of a pattern and practice of discrimination.28 Relying in

---

20. Id. at 483–84.
21. Id. at 489 n.7.
22. Id. at 496 n.17.
23. Id.

The “standard deviation” is a unit of measurement that allows statisticians to measure all types of disparities in common terms. Technically, a “standard deviation” is defined as “a measure of spread, dispersion, or variability of a group of numbers equal to the square root of the variance of that group of numbers.”

Palmer, 815 F.2d at 92 n.7 (quoting D. BALDUS & J. COLE, STATISTICAL PROOF OF DISCRIMINATION 359 (1980)). Case law often erroneously interchanges this term with the more technically appropriate term “standard error,” which describes the distribution of sample estimators, such as the mean, around its true value. See David H. Kaye & David A. Freedman, Reference Guide on Statistics, Federal Judicial Center, Reference Manual on Scientific Evidence 174 (2d ed. 2000).

26. Id. at 301.
27. Id. at 305.
28. Id. at 309.
part upon this finding, the Court remanded the case with instructions that the district court craft an acceptable remedy, which was to include injunctive as well as other equitable relief.\textsuperscript{29}

The Court’s reference in these cases to the scientific convention of “two or three” standard deviations, strictly speaking, is mere dicta. Because the disparities in each case were so great, the Court was not required to decide whether two or three, rather than say four or five, standard deviations should be the deciding criterion. As an analogy, consider a court that is asked to determine as a matter of law whether a man who is six-foot five and weighs six hundred pounds is morbidly obese. While it may be helpful for the court in affirmatively answering that question to observe that most men of that height weigh between two and three hundred pounds, that fact is not dispositive and the court readily could have reached that conclusion without reference to those data.

Thus, in \textit{Watson v. Fort Worth Bank & Trust}, Justice O’Connor reviewed the Court’s statistical criteria in employment discrimination cases.\textsuperscript{30} Although she acknowledged the prevalence of the Castaneda-Hazelwood test of “two or three standard deviations,” she emphasized that the Court never instructed lower courts to apply the standard mechanistically.\textsuperscript{31} Rather, courts should evaluate statistical evidence in relation to the disputed issues and determine the appropriateness of such evidence case by case.\textsuperscript{32} Justice O’Connor observed:

\begin{quote}
We have emphasized the useful role that statistical methods can have in Title VII cases, but we have not suggested that any particular number of “standard deviations” can determine whether a plaintiff has made out a prima facie case in the complex area of employment discrimination. Nor has a consensus developed around any alternative mathematical standard. Instead, courts appear generally to have judged the “significance” or “substantiality” of numerical disparities on a case-by-case basis. At least at this stage of the law’s development, we believe that such a case-by-case approach properly reflects our recognition that statistics “come in infinite variety and . . . their usefulness depends on all of the surrounding facts and circumstances.”\textsuperscript{33}
\end{quote}

Nevertheless, many lower courts have adopted the Castaneda-Hazelwood criterion of “two or three standard deviations” as a bright-line rule. In so doing, they often have noted that this criterion, when applied to the commonly assumed bell-shaped, normal distribution, corresponds to the .05 level of “statistical significance” prevalent in

\begin{itemize}
\item\textsuperscript{29} \textit{Id.} at 312–13.
\item\textsuperscript{30} \textit{487 U.S. 977} (1988).
\item\textsuperscript{31} \textit{Id.} at 995 n.3.
\item\textsuperscript{32} \textit{Id.}
\item\textsuperscript{33} \textit{Id.} (citations omitted).
\end{itemize}
the scientific literature. This criterion—the five percent probability threshold—corresponds, in turn, to the probability of “Type I error,” the probability of mistakenly rejecting the null hypothesis when it is true. Generally, the lower the probability of Type I error, the more confident the researcher is that he or she is not mistakenly claiming a statistical finding to be important. The Seventh Circuit has explained:

In addition to describing statistical significance in terms of levels of standard deviation, statistical significance also may be expressed as a probability value (P) on a continuous or relative scale ranging from 0 to 1.0. The level of statistical significance rises as the value of the (P) level declines. . . . A (P) value below .05 is generally considered to be statistically significant, i.e., when there is less than a 5% probability that the disparity was due to chance. For large samples, statistical significance at a level in the range below 0.05 or 0.01 is “essentially equivalent” to significance at the 2 or 3 standard deviation level.

Similar reasoning has led various courts to adopt a per se rule that statistical evidence failing to meet the .05 level of significance is inadmissible.

For example, in Palmer v. Shultz, the District of Columbia Circuit extensively considered the rather esoteric question of whether it should apply a one-tailed or two-tailed test of statistical significance. Its decision to apply the two-tailed test ultimately was outcome-determinative and led to a rejection of the plaintiff’s statistical evidence. Similarly, in Bennett v. Total Minatome Corp, the Fifth Circuit reaffirmed its rule that only “statistically significant” results, i.e., those exceeding two standard deviations, are admissible. In the same vein, the Eleventh Circuit has opined:

The “general rule” is that the disparity must be “greater than two or three standard deviations” before it can be inferred that the employer has engaged in illegal discrimination under Title VII. The Court has also called that sort of imbalance a “gross statistical disparity.”

34. In Castaneda, the Supreme Court noted that, when dealing with large numbers, social scientists reject the “hypothesis of equality”—that the chances of an event are “equally” likely to result from chance or a proposed cause—if a disparity between actual and expected representation exceeds two or three standard deviations. 430 U.S. at 496 n.17.

35. Griffin v. Bd. of Regents of Regency Univs., 795 F.2d 1281, 1291 n.19 (7th Cir. 1986) (citing Coates v. Johnson & Johnson, 756 F.2d 524, 537 n.13 (7th Cir. 1985)).

36. See, e.g., Bennett v. Total Minatome Corp., 138 F.3d 1053, 1062 (5th Cir. 1998).

37. 815 F.2d 84 (D.C. Cir. 1987).

38. Id. at 94–95.

39. Total Minatome, 138 F.3d at 1062.

Of course, not all authorities agree with “statistical significance” as the bright-line rule for admissibility. Judge Posner has suggested that the scientific convention of assigning “significance” only to statistical results characterized by a p-value of five percent or less reflects the space limitations of scientific journals, and the consequent need to publish papers selectively, rather than any scientific imperative.\footnote{Kadas v. MCI Systemhouse Corp., 255 F.3d 359, 362 (7th Cir. 2001).} The EEOC’s Uniform Guidelines on Employee Selection suggest a rule of thumb that should be applied in evaluating an employer’s selection procedures: a protected group selection rate that is less than eighty percent of a preferred group selection rate warrants closer scrutiny of the selection procedures involved.\footnote{29 C.F.R. § 1607.4(D).}

Few courts have embraced the eighty percent rule, in part because it potentially contradicts the Supreme Court’s reliance on the “two to three standard deviations” criterion.\footnote{See, e.g., Clady v. County of Los Angeles, 770 F.2d 1421, 1428 (9th Cir. 1985) (rejecting the eighty percent test), cert. denied, 475 U.S. 1109 (1986).} For example, in cases involving small numbers of plaintiffs, disparities that are too small to be statistically significant, and thus of no probative value in the eyes of some courts, nevertheless may fail the eighty percent test.\footnote{See, e.g., Mems v. City of St. Paul, 224 F.3d 735, 740–41 (8th Cir. 2000) (refusing to apply eighty percent rule to sample size too small to be statistically significant); Nash v. Consol. City of Jacksonville, 895 F. Supp. 1536, 1542–43 (M.D. Fla. 1995), aff’d mem., 85 F.3d 643 (11th Cir. 1996). But see Pietras v. Bd. of Fire Comm’rs, 180 F.3d 468, 474–75 (2d Cir. 1999) (holding that sample size of seven was not too small for application of eighty percent rule when corroborated with other evidence of gender bias).} Similarly, a disparity that is within the eighty percent range nevertheless may be statistically significant because it exceeds two or three standard deviations. A recent decision by a Sixth Circuit panel illustrates this ambiguity; although the panel majority held that a statistically significant disparity was probative of discrimination, a dissenting opinion advocated that the evidence was unpersuasive because the eighty percent rule was the circuit’s one and only criterion regarding the admissibility of statistical evidence.\footnote{Isabel v. City of Memphis, 404 F.3d 404, 416–18 (6th Cir. 2005) (Batchelder, J., dissenting).}

IV. Two Problems with Using Statistical Significance or Standard Deviations as the Benchmark in Pattern and Practice Cases

There are two categories of problems with using statistical significance as the benchmark in pattern and practice cases, namely, the problems caused by (1) variations in sample size and (2) inconsistent treatment of statistical evidence depending on the theory of discrimination pursued.
A. Variations in Sample Size

Commentators have sought to distinguish mere statistical significance from the less well-defined concepts of “practical significance” or “legal significance” because the materiality of statistically significant results can be illusory. For example, other things being equal, as the size of the sample increases, or in discrimination cases, as the number of employees or employment decisions increases, so do the number of standard deviations associated with any given disparity. As a result, numerical disparities that would be considered de minimis by any reasonable standard may rise to the level of statistical significance.

Professor Daniel Rubinfeld has observed:

Practical significance means that the magnitude of the effect being studied is not de minimis—it is sufficiently important substantively for the court to be concerned. For example, if the average wage rate is $10.00 per hour, a wage differential between men and women of $0.10 per hour is likely to be deemed practically insignificant because the differential represents only 1% ($0.10/$10.00) of the average wage rate. That same difference could be statistically significant, however, if a sufficiently large sample of men and women was studied. The reason is that statistical significance is determined, in part, by the number of observations in the data set. . . . Thus, a $1 per hour wage differential between men and women that was determined to be insignificantly different from zero with a sample of 20 men and women could be highly significant if the sample were increased to 200.46

This principle is illustrated by contrasting two prominent cases. In *Castaneda*, the Supreme Court set forth the formula for calculating standard deviations.47 Using that formula, the Court found that the nearly 100% disparity between the actual and expected numbers of Mexican-American grand jurors was meaningful, exceeding twenty-nine standard deviations.48 However, instead of asking whether the disparity between 688 and 339 is significant, consider the related question of how small a disparity would satisfy the two or three standard deviation criterion? If only 655 or fewer Mexican Americans had been selected, this would barely amount to two standard deviations. Thus, a disparity of 25/688, or four percent, is the minimum disparity necessary for the statistical comparison to be probative according to that criterion.

Next, consider the extreme, but nevertheless real, example presented in *Dukes v. Wal-Mart Stores, Inc.*49 As the nation’s largest private employer, Wal-Mart makes tens of thousand of promotion decisions each year. For example, the plaintiffs’ expert in *Dukes* identified nearly

---

47. 430 U.S. at 496 n.19.
48. Id.
49. 222 F.R.D. 137 (N.D. Cal. 2004).
50,000 promotions to the management ranks alone between 1997 and 2002.\textsuperscript{50} This expert’s data also suggests that female employees constituted roughly sixty percent of the group eligible for these promotions.\textsuperscript{51} Thus, the proportional or “expected” number of females promoted would be 30,000 (50,000 x .60). If the \textit{Castaneda} formula is applied to these factual contentions, a two standard deviation disparity would result if Wal-Mart promoted 29,780 women instead of the expected 30,000, because the standard deviation is approximately 110.\textsuperscript{52} This disparity of 220, which would equal two standard deviations, is just seven-tenths of one percent (.007) of the expected number (30,000), a fraction of the four percent threshold implicit in \textit{Castaneda}, and far less than most would consider a “gross” disparity.

This property of “statistical significance” as a yardstick that shrinks the required disparity (percentage-wise) as the number of employment decisions grows was of negligible importance when courts initially searched for a threshold measure of admissibility. The earliest cases concerned relatively small populations, and the numerical disparities before the courts often were large enough to be eye-catching without resort to formal statistical measures, e.g., the “inexorable zero” in \textit{Teamsters}, and the near-halving of the expected number of grand jury summonses to Mexican Americans in \textit{Castaneda}.\textsuperscript{53} In contrast, employment class actions now may involve millions of employee records, which today’s computers manage easily. Indeed, the putative class in \textit{Dukes} is alleged to consist of 1.5 million present and former female employees.\textsuperscript{54} Moreover, the employment decisions at issue involve not just these class members but perhaps an even larger number of comparators (non–class members), who also must be included in any analysis. The result is that disparities between groups that previously were deemed negligible, because much smaller samples were studied, now may be judged “statistically significant” and, according to many courts, probative of discrimination.

“Statistical significance” is shorthand for a researcher’s conclusions with respect to a specific null hypothesis—typically, that rates

\begin{itemize}
\item \textsuperscript{51} \textit{Id.}
\item \textsuperscript{52} \textit{Castaneda}, 433 U.S. at 496 n.17.
\item \textsuperscript{53} Note that the number of decisions at issue in \textit{Teamsters} was 1,828, in \textit{Castaneda}, 870; and in \textit{Hazelwood}, 405. \textit{See Teamsters}, 431 U.S. at 337; \textit{Castaneda}, 430 U.S. at 496 n.17; \textit{Hazelwood}, 433 U.S. at 310.
\item \textsuperscript{54} \textit{Dukes}, 222 F.R.D. at 142.
\end{itemize}
of pay, promotion, hiring, layoff, etc., between two demographic groups are identical. When the data are clear enough to permit the researcher to discern some difference between the two groups, with a specified degree of accuracy, the disparity is statistically significant. In contrast, when the researcher cannot reject the null hypothesis by that standard, the disparity may be “statistically insignificant.” However, the ability of any researcher to discern a disparity depends not only upon the magnitude of the difference but also the clarity with which the data permit any difference to be discerned.

Increasing the sample size, therefore, is equivalent to arming a researcher with a high-powered telescope. Disparities that the researcher judged too small to discern confidently in smaller samples can be measured with confidence in larger samples. In fact, miniscule disparities in very large samples can be statistically significant, as the Wal-Mart example indicates, while much larger disparities can be rejected as insignificant in much smaller samples. In sum, statistical significance merely addresses whether there is probative evidence of some disparity but is not informative as to whether that disparity is “gross,” as that term is conventionally used.

B. Inconsistent Treatment of Statistical Evidence Depending on the Theory of Discrimination Pursued

Besides being ambiguous, “statistical significance” poses a conceptual problem in terms of how courts treat evidence under alternative theories of discrimination that meet this threshold. Although statistical evidence is ubiquitous in pattern and practice class actions, courts also admit statistical evidence in garden-variety, single-plaintiff discrimination cases. In these cases, statistical evidence may assist the plaintiff in establishing a prima facie case of discrimination. If admissible, such evidence shifts the burden of production, though not persuasion, to the employer to articulate a legitimate, nondiscriminatory reason for its challenged action. Once the employer provides a reason for its decision that is specific to the plaintiff, generalized statistical evidence no longer is probative and a plaintiff cannot prove “pretext” by relying solely on that evidence. That is not to say that statistical proof can never be probative at this stage, but rather that the plaintiff

55. “A nondiscriminatory practice that produced a very small difference in the results for two different groups would eventually become statistically significant if the sample size, and therefore the power [of the test], were to increase.” Stephen E. Fienberg, Samuel H. Krislov & Miron L. Straf, Understanding and Evaluating Statistical Evidence in Litigation, 36 JURIMETRICS J. 1, 23 (1995).


58. Barnes, 896 F.2d at 1468–69.
must adduce evidence that speaks to the specific explanation proffered by the employer. At no point in a single-plaintiff, disparate-treatment lawsuit does the statistical evidence, no matter how “gross,” shift the burden of persuasion to the employer.

In contrast, gross statistical disparities suffice to shift the burden to the employer in pattern and practice (class action) cases, which potentially affect the claims of perhaps tens of thousands of employees. Although the statistical evidence in these cases often is supported by anecdotal evidence that brings these statistics “convincingly to life,” these anecdotes, by definition, pertain directly to only a tiny portion of the class. Yet, the presumption created primarily by this statistical proof applies to each and every class member and requires the employer to rebut that presumption in each specific instance.

The Supreme Court’s Teamsters decision explains that, in the liability phase of a pattern and practice case, “the focus often will not be on individual hiring decisions, but on a pattern of discriminatory decision-making.” However, the pattern and practice, once proved, must be sufficiently pervasive to “create a greater likelihood that any single decision was a component of the overall pattern.” Presumably, it is because statistical evidence in a pattern and practice case must be strong enough to shift the burden of proof, whereas this strong showing is not required of a single plaintiff, that the court insists these disparities must be “gross.”

Although lower courts recognize as well that proof of a pattern and practice of discrimination requires evidence of gross statistical disparities, few courts, if any, have considered whether the “two or three standard deviations” referenced in Castaneda suffices in a particular case to create a greater likelihood that the decisions affecting each class member were a component of the overall pattern. Case law regarding this critical issue, however, is much further developed in other legal contexts.

V. The Statistical Proof Required to Create a Presumption That a Wrongful Act Was the Cause-in-Fact of an Injury

Daubert v. Merrell Dow Pharmaceuticals, Inc., the seminal case on the admissibility of expert testimony, also illustrates how courts

60. Franks, 424 U.S. at 772.
61. Teamsters, 431 U.S. at 360 n.46.
62. Id. at 359 n.45.
63. Bell v. Envtl. Prot. Agency, 232 F.3d 546, 553–54 (7th Cir. 2000) (stating statistical evidence may be “relevant to and probative of the issue of pretext even when it is insufficient to support a pattern and practice disparate treatment case” and “the evidence that blacks are not promoted as often as nonblacks, even though not statistically significant, is still circumstantial evidence of possible discrimination”).
64. 430 U.S. at 496 n.17.
create presumptions in tort law based upon statistical proof. The substantive question posed in Daubert was whether the drug Bendectin, administered to pregnant women to ease morning sickness, is a teratogen (a substance that causes limb reduction birth defects). The plaintiffs proffered expert testimony on this question, which the defendant moved to strike, contending that these opinions were “unscientific.”

Although the Supreme Court famously addressed the standard for admitting expert testimony, it remanded the case to the Ninth Circuit to determine the threshold level of proof sufficient to prove that Bendectin was the cause-in-fact of the plaintiffs’ birth defects. The parties did not dispute that the plaintiffs’ mothers had ingested Bendectin during pregnancy; rather, causation was complicated by the fact that, regardless of whether a mother ingests Bendectin, she faces a risk of approximately 2–3% of having a child with a limb reduction defect. The difficult question, therefore, was how to determine whether any child born with a birth defect more likely than not suffered that defect because of Bendectin, given that most birth defects occur for unknown reasons.

The Daubert plaintiffs relied principally on epidemiological evidence to prove causation. Their experts testified to the enhanced incidence of birth defects among children exposed to Bendectin relative to those who were not exposed. This evidence, in and of itself, however, was not sufficient to establish causation for reasons unrelated to statistical significance. As the Ninth Circuit observed, California tort law requires that the plaintiffs “show not merely that Bendectin increased the likelihood of injury, but that it more likely than not caused their injuries.” This, of course, is the familiar cause-in-fact requirement that applies not only at common law, but under certain antidiscrimination laws as well.

66. Id. at 582.
67. Id. at 582–83.
68. Daubert v. Merrell Dow Pharm., Inc., 43 F.3d 1311 (9th Cir. 1995).
69. Id. at 1313.
70. Id. at 1313–14.
71. Id. at 1314.
72. Id.
73. Id. at 1321–22.
74. Id. at 1320.
75. See, e.g., Hartman v. Moore, 126 S. Ct. 1695, 1701 (2006) (“we have held that retaliation is subject to recovery as the but-for cause of official action offending the Constitution”); Pineda v. United Parcel Serv., Inc., 360 F.3d 483, 487 (5th Cir. 2004) (holding that plaintiff must establish “but for” causation under Title VII retaliation provision); Horwitz v. Bd. of Educ. of Avoca Sch. Dist. No. 37, 260 F.3d 602, 610 (7th Cir. 2001) (holding “but for” causation standard applies to ADEA cases). See also Byers v. City of Albuquerque, 150 F.3d 1271, 1274–75 (10th Cir. 1998) (police officer lacked standing to challenge affirmative action policy because evidence established he would not have qualified for promotion notwithstanding the alleged effects of the policy); Constr. & Gen. Laborers’ Local Union No. 230 v. City of Hartford, 153 F. Supp. 2d 156, 161–62 (D. Conn. 2001) (contractor could not challenge minority set-aside because no evidence established the contractor would have been awarded the contract absent the set-aside program).
The Ninth Circuit held that a plaintiff who relies upon epidemiological evidence

must establish not just that their mother's ingestion of Bendectin increased somewhat the likelihood of birth defects, but that it more than doubled it—only then can it be said that Bendectin is more likely than not the source of their injury. Because the background rate of limb reduction defects is one per thousand births, plaintiffs must show that among children of mothers who took Bendectin the incidence of such defects was more than two per thousand.76

The court calculated the incidence of a birth defect as the ratio of the number of births in a given period that manifest that defect relative to the total number of births.77 Thus, if a population gives birth to 100,000 children, 500 of whom suffer from the defect, the incidence is .005. Now consider the incidence of birth defects among a second population that has been exposed to the drug. Suppose this population also gives birth to 100,000 children, 1,000 of whom suffer from similar birth defects. The incidence of the defect among this population is .010. Epidemiologists speak in terms of the “relative risk,” the ratio of the defect in the exposed population relative to its incidence in the control population. In the foregoing example, the incidence ratio is 2.0 because the exposed population had an incidence of .010, which is divided by the incidence among the control population, which is .005.

The rationale for setting the minimum relative risk at 2.0 is the following. For a plaintiff to prevail against a drug manufacturer, the plaintiff must prove that the drug more likely than not caused the plaintiff’s injury. If the 1,000 children in the exposed population who suffered birth defects are a group, then the probability is that 500 members of this group would have suffered the same birth defect notwithstanding their exposure to the drug, just as in the control population. Analogously, there is another group of 500 among the 1,000 children who suffered a birth defect solely because of their exposure to the drug. Thus, for any particular plaintiff among the class of 1,000, the chances are 50-50 that this individual would have had the same birth defect notwithstanding exposure to the drug. Because the standard of proof is “more likely than not,” only if the relative risk exceeds 2.0 is it more likely than not that any given plaintiff was injured by the drug.78

At issue here is the distinction between proving that a wrong exists and proving that it injured a particular plaintiff; the tort law distinguishes these concepts in terms of “general” and “specific” causation.79

76. Daubert, 43 F.3d at 1320.
77. Id. at 1321.
“General causation is whether a substance is capable of causing a particular injury or condition in the general population, while specific causation is whether a substance caused a particular individual’s injury.”

In employment discrimination, this is equivalent to distinguishing between whether discrimination may have tainted an employer’s promotional selections in general, as opposed to whether the same employer would have promoted a particular plaintiff absent the discriminatory motives. Statistical proof that the relative risk is greater than 1.0 suffices to establish the former; however, a plaintiff must prove that it is greater than 2.0 to establish specific causation. Because evidence in a pattern and practice case must be sufficient to establish a presumption that each class member suffered discrimination, as opposed to no class member in particular, the 2.0 criterion should establish the minimum disparity that shifts the burden of proof.

VI. Why Do Statistical Standards Appear to Differ Between Tort and Employment Law?

An obvious explanation of why courts have applied different thresholds to statistical evidence in tort and employment cases emphasizes the different issues to which the evidence is addressed in each instance. In the Bendectin cases, a plaintiff suffering a birth defect typically has the burden of proving that there is a greater than fifty percent probability that the drug caused his or her defect. As in all tort cases, a plaintiff cannot prevail merely by establishing that conduct was generally wrongful without also establishing that this wrong was the cause-in-fact of the plaintiff’s injury. Therefore, courts routinely require plaintiffs who rely primarily on epidemiological evidence to prove that the relative risk is greater than 2.0, holding that statistical evidence below this threshold fails to establish specific causation.

In contrast to tort claims, employment law statutes provide a variety of remedies and also may impose different burdens of proof, depending upon the claims at issue. For example, plaintiffs often seek injunctive relief in order to halt ongoing discrimination. Whether an injunction is appropriate depends, in part, on proof that the discriminatory policy exists, rather than how it has affected any particular employee. Consequently, in the language of tort law, proof of general causation, along with evidence satisfying the other requisites, may entitle a plaintiff to an injunction. In statistical terms, proof of a relative risk significantly

81. See Allison, 184 F.3d at 1315 n.16; Smith, 278 F. Supp. 2d at 691 n.9.
82. Teamsters, 431 U.S. at 360 ("At the initial, 'liability' stage of a pattern-or-practice suit the Government is not required to offer evidence that each person for whom it will ultimately seek relief was a victim of the employer's discriminatory policy. Its burden is to establish a prima facie case that such a policy existed.").
greater than 1.0 may satisfy that burden. This criterion is reflected, at least implicitly, in the many employment law decisions that reference statistical significance as the guiding principle for determining legal sufficiency or deciding the admissibility of expert testimony.83

_Castaneda_, the progenitor of statistical cases in employment law, demonstrates why this minimal standard applies.84 The claim in _Castaneda_ was that Mexican Americans systematically were excluded from juries.85 That allegation was made not by a potential juror, seeking to establish that he or she was excluded on prohibited grounds, but by a criminal defendant seeking to overturn the indictment of an allegedly tainted grand jury.86 Thus, it was not the criminal defendant's burden to prove that any particular Mexican American was excluded from the jury, but merely that the practice excluded Mexican Americans in general. Accordingly, the Court appropriately noted that the disparity in selection rates exceeded “two or three standard deviations”—the threshold standard associated with proof of general causation—in determining that the criminal defendant had adduced sufficient evidence to prove his case.87

Similarly, in _Teamsters_ and _Hazelwood_, the seminal pattern and practice cases, the government sought to vindicate the rights of employees protected by Title VII.88 The government’s burden was not to prove that it was injured by discrimination. Rather, section 707 of the Civil Rights Act empowers the government to bring a pattern and practice suit on behalf of persons injured by violations of Title VII.89 The government’s burden in these suits is to establish the existence of the pattern and practice. Although it may present anecdotal evidence regarding individual employees, _Hazelwood_ suggests that the government may prove the pattern and practice by statistics alone.90 Indeed, the Court’s reference to the criterion of “two or three standard deviations” suggests that proof of a relative risk greater than 1.0 might be sufficient to carry this burden.91

Since these cases were decided, Congress has passed the Civil Rights Act of 1991.92 As a result, many discrimination plaintiffs seek not merely to enjoin a discriminatory practice, but to obtain individual monetary relief as well. This development is amply reflected in decisions regarding class certification. Prior to the 1991 Act, Title VII cases

83. See, e.g., _Anderson_, 180 F.3d at 341.
84. _Castaneda_, 430 U.S. 482.
85. _Id._ at 483–84.
86. _Id._ at 482.
87. _Id._ at 496 n.17.
89. 42 U.S.C. § 2000e-6(a).
90. _Hazelwood_, 433 U.S. at 307–08.
91. _Id._ at 311 n.17.
proceeded largely, if not exclusively, under Rule 23(b)(2), which permits classes to be certified when injunctive relief is the primary objective.\textsuperscript{93} Presently, with monetary relief a substantial objective of employment class actions, a hotly disputed question is whether plaintiffs can continue to certify these classes under Rule 23(b)(2) or whether they must seek certification under Rule 23(b)(3), which is a more exacting standard. Indeed, other circuits have followed \textit{Allison v. Citgo Petroleum Corp.} in establishing virtually a bright-line rule that plaintiffs seeking monetary relief, even backpay, generally must proceed under Rule 23(b)(3).\textsuperscript{94}

Suits for monetary relief, as opposed to injunctive relief, generally require proof of individual injury. As a result, plaintiffs typically must demonstrate, under many statutes and the Constitution, that the unlawful conduct they allege was the “but for” cause of their personal injury, or at least a motivating factor.\textsuperscript{95} The distinction between the burden statistical evidence must carry when the objective is monetary relief as opposed to injunctive relief may be gleaned from a recent decision in an employment case by Judge Easterbrook of the Seventh Circuit:

\begin{quote}
A statistical analysis may be able to answer these questions—and, if the answer is yes, the knowledge that high dosage x-rays increase the risk of cancer may inform a decision whether the benefits of the procedure are worth the extra risk. But it will not tell us whether a given person who develops cancer did so because of the x-ray; only 2.5% of cancers can be attributed to the radiation, so 97.5% of all cancers, even among persons exposed to high-dosage x-rays, have other causes. This is the sense in which statistics are more helpful in a pattern-or-practice case, where a judge will be asked to direct the employer to change how it makes hiring or promotion decisions.\textsuperscript{96}
\end{quote}

Although the Supreme Court recognized in \textit{Desert Palace} that the 1991 Amendments to the Civil Rights Act require a plaintiff to prove merely that discrimination was the “motivating factor” behind an adverse action, backpay and front pay remain appropriate remedies only for those plaintiffs who in fact suffered an economic injury because of discrimination.\textsuperscript{97} Because the statute’s antiretaliation provision was

\begin{footnotes}
\footnotetext[96]{Baylie v. Fed. Reserve Bank of Chicago, 476 F.3d 522, 524 (7th Cir. 2007) (emphasis added).}
\footnotetext[97]{\textit{Desert Palace}, 539 U.S. at 101.}
\end{footnotes}
not amended in that fashion, the “cause in fact” requirement continues to apply in those cases.98

Accordingly, there is no inherent difference in the statistical standards that apply to tort claims and discrimination claims. Rather, to the extent differences exist, they should reflect the burden of proof assigned to plaintiffs in each category of cases, which in turn reflects the particular relief sought. Indeed, discrimination claims that seek monetary relief now may more closely resemble common law tort claims than discrimination claims of an earlier era, in which injunctive relief was the primary goal.

VII. Can a Relative Risk of 2.0 Be the Prevailing Statistical Standard in Pattern and Practice and Other Employment Cases?

Plaintiffs in a pattern and practice case, the theory commonly alleged in class action discrimination suits, seek to prove that discrimination is an employer’s standard operating procedure, and thereby to create a presumption that each member of the class is a victim of that discrimination.99 In this respect, these plaintiffs are undertaking the same burden as class representatives in tort cases, who allege that their injuries, along with those of the class, more likely than not were caused by exposure to the challenged product or practice. If they succeed, each group of plaintiffs has proved its entitlement to individual relief unless the employer can demonstrate that the injury to any particular plaintiff had a more benign cause.100

The Supreme Court explained in Teamsters that

[pro]sumptions shifting the burden of proof are often created to reflect judicial evaluations of probabilities and to conform with a party’s superior access to the proof. These factors were present in Franks. Although the prima facie case did not conclusively demonstrate that all of the employer’s decisions were part of the proved discriminatory pattern and practice, it did create a greater likelihood that any single decision was a component of the overall pattern.101

This presumption of a “greater likelihood” does not follow from proof of mere statistical significance, which establishes a relative risk greater than 1.0.102 This relative risk is insufficient to create a presumption

---

100. Id.
101. Id. at 359 n.45 (citations omitted).
102. Because statistical significance in discrimination cases typically is expressed with respect to the null hypothesis of equality in the treatment of two groups, it bears only on whether the relative risk is 1.0, because the equal incidence of termination, layoff, or hiring refusals between two groups, when expressed as a ratio, equals 1.0.
regarding the “greater likelihood” that the decision regarding any class member stems from a pattern and practice, because it does not explicitly account for nondiscriminates to whom the employer similarly denied promotions or job offers for reasons obviously unrelated to a protected characteristic.

Returning to Dukes, the plaintiffs’ expert opined that data indicating that 41,000 incumbent employees applied for the store manager position between 1997 and 2002 may understate the actual number. Further, he identified 3,567 applicants whom Wal-Mart promoted to that position during the period under study, eighteen percent of whom were women. Because women constituted twenty-two percent of the pool of employees eligible for promotion to store manager, this expert discerned a “shortfall” of 155.7 in the promotion of women, which he opined was statistically significant, with a “z-value” of 7.72. Although the plaintiffs advanced this evidence in support of a pattern and practice theory, the relative risk implied by that analysis is well below 2.0.

The vast majority of applicants for the store manager position, more than 37,000 men and women, or 91.3 percent, were unsuccessful. Any assistant or co-managers who applied for that position—the jobs the expert determined to be the “availability pools”—had but an 8.7 percent chance of being selected, irrespective of gender. Indeed, 29,000 males, in addition to 8,390 females, were rejected for promotion. If Wal-Mart had promoted women proportionally, only 8,235 would have been rejected—because women constituted twenty-two percent of the pool of 41,000, which is characterized by a 91.3 percent rejection rate overall. The gender bias this expert found in these statistics increased the rejection rate for women in the availability pool from 91.7 percent to ninety-three percent. The relative risk associated with alleged gender bias is just 1.01—large enough to be statistically significant, but by no means sufficient to raise a presumption that each rejected female applicant suffered discrimination. Yet the district court, in which plaintiffs sought to certify a class of females complaining of a pattern and practice of gender bias in promotions, concluded that “Dr. Drogin’s method is sufficient to create an inference of discrimination for purposes of this motion.”

103. See Drogin, supra note 50, at 29–31.
104. Id. at 36 (Table 25).
105. Id. A z-value is roughly equivalent to the number of standard deviations. “The Z-Score measures the number of standard deviations by which a particular observed outcome differs from its expected value, assuming random selections from the relevant pool.” Joseph L. Gastwirth, Employment Discrimination: A Statistician’s Look at Analysis of Disparate Impact Claims, 11 LAW & INEQ. 151, 154 n.14 (1992).
106. See Drogin, supra note 50, at 35.
VIII. A Standard of Less Than 2.0 May Violate the Rules Enabling Act

The Rules Enabling Act, 28 U.S.C. § 2072(b), prohibits a rule of procedure from “abridg[ing], enlarg[ing] or modify[ing] any substantive right.” Ordinarily, a Title VII plaintiff cannot recover backpay merely by proving a diminished likelihood of being hired, promoted, etc., because of discrimination. To the contrary, an individual plaintiff must establish that any adverse employment action he or she suffered was because of discrimination. Yet, in class actions, backpay routinely is divided among class members based upon a relative risk greater than 1.0, despite the absence of proof that any particular class member was denied any specific employment opportunity because of discrimination, and strong evidence that the vast majority of class members certainly were not. For example, in scenarios such as the Dukes illustration, it is commonplace for courts to value in the aggregate the 155 lost promotional opportunities for women, estimated by the plaintiffs’ expert, among the roughly 8,400 class members. Yet none of these class members could prove her entitlement to backpay, were she to sue individually, based solely upon evidence that gender discrimination reduced her chances of promotion from 8.3 to seven percent.

Indeed, the Dukes court was undeterred by this apparent inconsistency:

The cases acknowledge that awarding backpay to all potential victims of the employer’s policy (as opposed to just the actual victims) has the effect of generating a “windfall for some employees who would have never been promoted had vacancies been filled on a [non-discriminatory] basis and undercompensating the genuine victims of discrimination.”


109. Drogin, supra note 50, at 36.

110. For example, the Fifth Circuit has stated that where multiple job applicants failed to secure a position, “only those who can prove that they would have gotten the position but for the discrimination can recover compensatory damages.” Arnold v. U.S. Dept of the Interior, 213 F.3d 193, 197 (5th Cir. 2000). In Bishop v. Gainer, 272 F.3d 1009, 1016 (7th Cir. 2001), the court analogized situations in which just one of several plaintiffs could have received a contested promotion to the problem addressed by the tort doctrine of “lost-chance of survival.” Under that theory, a medical malpractice plaintiff who had only a twenty-five percent chance of survival if a procedure were performed properly could recover only twenty-five percent of the total damages if the procedure is performed negligently. In employment cases, the analogy would be that a plaintiff with just a twenty-five percent chance of being selected in the absence of discrimination, could recover no more than that percentage of backpay. However, several states apply the “lost-chance” doctrine only if the chance of survival absent any negligence is greater than fifty percent. See, e.g., Kramer v. Lewisville Mem’l Hosp., 858 S.W.2d 397, 400–01 (Tex. 1993) (collecting cases that recognize and fail to recognize the doctrine); see also Baylie, 476 F.3d at 525 (suggesting that an employee who suffered a diminished opportunity of a promotion because of discrimination could claim damages determined by the percentage reduction in the likelihood of promotion). However, research has found no case in which a court has approved an individual plaintiff’s recovery on that basis.
discrimination by forcing them to share the award with their undeserving brethren." For example, in *O & G Spring & Wire Forms*, backpay for 17 positions that would have gone to class members absent discrimination was spread over 450 persons who identified themselves as having applied (or having been deterred from applying) for the positions at issue; in *Hameed*, backpay for 45 apprentice positions was spread over a much larger group of "potential discriminates . . . ."

The cases conclude, however, that this "rough justice" is better than the alternative of no remedy at all for any class member.111

However, these cases do not come to grips with the assumption in *Segar v. Smith* that the relative proportion of "deserving" and "nondeserving" class members may be reversed when a relative risk of 1.0 defines the threshold level of proof.112 In the promotion example drawn from *Dukes*, although the "shortfall" in female promotions was estimated to be just 155, payment would be made to 8,080 "nondeserving" class members (the 8,235 rejected applicants minus the 155 who would have been promoted).113 This is hardly "rough justice" but rather the payment of windfalls to the overwhelming majority of class members who, even in a nondiscriminatory world, still would not have been promoted. Thus, in this instance, the class action device serves to increase the backpay recovery of class members, and the financial liability of the employer (because no court has awarded backpay to an individual plaintiff who merely proves a reduced probability of promotion of 1.3 percentage points), in seeming contravention of the Rules Enabling Act.114

**IX. Conclusion**

This paper proposes simply that the standard for determining the sufficiency of statistical evidence should depend on the particular issue to which that evidence is addressed. Because courts can enjoin discrimination even though its particular victims are identifiable only as a class, a relative risk greater than 1.0, or statistical significance, is a

---

111. *Dukes*, 222 F.R.D. at 177 (citations omitted). *See also* Thomas v. City of Evanston, 610 F. Supp. 422, 435–36 (N.D. Ill. 1985) (in Title VII case, class-wide backpay award based on statistical computations was appropriate despite ignorance as to which class members would have been hired); E.E.O.C. v. Chicago Miniature Lamp Works, 640 F. Supp. 1291, 1298–1300 (N.D. Ill. 1986) (in Title VII case, individual hearings not necessary to determine entitlement to backpay; classwide award would be determined on statistical basis), rev'd on other grounds, 947 F.2d 292 (7th Cir. 1991).

112. 738 F.2d 1249, 1291 (D.C. Cir. 1984).

Though [Title VII] generally does not allow for backpay to those whom discrimination has not injured, this section should not be read as requiring effective denial of backpay to the large numbers of agents whom [defendant's] discrimination has injured in order to account for the risk that a small number of undeserving individuals might receive backpay.

*Id.*


114. On the other hand, requiring a relative risk of at least 2.0 ensures that at least a majority of the payments to the class will go to actual victims of discrimination.
reasonable criterion by which to define the threshold level of proof. On the other hand, the efficiency of the pattern and practice framework, when the goal is individual monetary relief, inheres in the “greater likelihood” presumption. As the tort cases demonstrate, that presumption requires proof of a relative risk greater than 2.0. If the plaintiff cannot cross that threshold because the evidence is not strong enough to reject the hypothesis that it is 2.0 or less, the pattern and practice framework is inappropriate for adjudicating questions regarding monetary relief. In terms of class certification, a relative risk greater than 1.0 is consistent with certifying a (b)(2) class seeking injunctive relief, but not a (b)(3) class, in which monetary relief is the primary objective; a (b)(3) class must establish that the relative risk is greater than 2.0.

As a result, the usefulness of statistical evidence in cases certified under Rule 23(b)(3) may be quite circumscribed. Unlike tort cases, in which a small percentage of the general population experiences injuries such as birth defects, adverse employment actions such as the failure to be hired or promoted are commonplace. Many employers surely reject a much larger number of applicants than they hire, or promote just a small fraction of those employees who are eligible for promotion. Under these circumstances, a doubling of the failure rate, i.e., a relative risk of 2.0, is impossible. For example, if Wal-Mart hires only thirty percent of male applicants for a position, meaning it rejects seventy percent, female plaintiffs obviously cannot prove that Wal-Mart rejected 140 percent of their group, the showing necessary to establish a relative risk of 2.0.

Under these assumed facts, no statistical showing would suffice to establish the “greater likelihood” that discrimination caused any rejected applicant to be denied hiring or promotion. However, the same statistical evidence would help to establish the existence of the discriminatory practice and pave the way for injunctive relief that would end it—the issues relevant to a class certified under Rule 23(b)(2). Nor is statistical evidence the sole means of proving a pattern and practice of discrimination.115 Thus, merely because statistical evidence may not be helpful in proving any particular class member's entitlement to monetary relief does not mean that the pattern and practice framework, and proof of a relative risk greater than 1.0, should not remain a cornerstone of class action litigation.

---

115. Pitre v. W. Elec. Co., Inc., 843 F.2d 1262, 1267 (10th Cir. 1988) (“[a]lthough statistics may be useful to show differences in treatment and to establish a pattern and practice, they are clearly not required”).