High Court's Old, Bad Stats Analysis Can Miss Discrimination

By **Daniel Levy** (November 16, 2023)

Daniel Levy

On Sept. 13, the U.S. Equal Employment Opportunity Commission continued to expand its access to employer data for statistical analysis through a memorandum of understanding with the Wage and Hour Division of the U.S. Department of Labor, which describes the data-sharing protocols between the two agencies.[1]

This follows a 2022 Office of Federal Contractor Compliance Programs directive laying out eight statistical tests likely to be acceptable, potentially among others, in OFCCP pay discrimination reviews.[2]

This continued push toward more frequent and powerful statistical analysis in discrimination regulation and litigation brings us back to the fundamental question of what statistical evidence "indicates that the discrepancy is significant," as the U.S. Supreme Court put it in Castaneda v. Partida in 1977.[3]

What is the Hazelwood/Castaneda standard of evidence of discrimination?

"Is the difference more than two standard deviations?"

That is the first question most attorneys ask about the difference between the plaintiff demographic group and its similarly situated benchmark in employment and compensation discrimination cases, and it is the wrong question.

The question refers to the commonly held interpretation of the Supreme Court decision in Hazelwood School District v. United States in 1977.[4] In that case, the Supreme Court, citing its own opinion in Castaneda, a jury selection case from earlier that year, reiterated that "as a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations," the difference would be suspect.[5]

The Supreme Court was right that the statistical analysis needed to determine whether Mexican American jurors were underrepresented in Castaneda was the same type needed to analyze whether African American teachers were underrepresented in Hazelwood. It is through this citation in Hazelwood to the Supreme Court's Castaneda opinion that the two or three standard deviations of difference between the plaintiff group and the benchmark demographic group made its way into employment discrimination cases.

As we will see, the Supreme Court's statement in Castaneda, as repeated in Hazelwood, was accurate enough for the very large degree of disparity observed in those cases. However, for many other cases involving disparate treatment and pay disparity, the Supreme Court's statement about "two or three standard deviations" is at best so imprecise, and at worst simply so wrong, that it has led to confusion and inaccurate determinations about statistical evidence in courts handling disparate treatment and compensation cases.

What error did the Supreme Court make in defining statistical evidence?

The Supreme Court, of course, did not invent this statistical measure of "standard deviations" out of whole cloth. The court relied on academic works, citing three.[6] Those

works describe the underlying statistical concepts, demonstrating that as the proportion of African American teachers among all teachers hired by the Hazelwood School District fell below the regional proportion of African American teachers, it became increasingly unlikely that the shortfall resulted from random chance, providing prima facie evidence of discrimination.

When flipping a coin 100 times, it would not be surprising to find 48 heads and 52 tails, or vice versa, even if the coin is fair. Similarly, it also would not be surprising if the percentage of African American teachers hired in the school district fell a little below the percentage of African American teachers in the regional hiring pool. Although we expect 50 heads and 50 tails, there is some degree of deviation from that expectation that should not surprise anyone.

Statisticians have figured out how unlikely an outcome is based on how different it is from the expected outcome in the absence of bias. With 100 flips of a fair coin, there would be only a 5% chance of getting fewer than 40 heads or more than 60 heads, and there would be a little more than a 1% chance of seeing fewer than 37 heads or more than 63 heads.

The same set of underlying calculations can be used to determine the likelihood of seeing any result drawn from some population, such as the number of African American teachers hired by the Hazelwood School District, based on the fact that at the time of Hazelwood about 5.7% of the teachers in the relevant region were African American.[7]

"The significance levels in most common use are 5%, 1% and 10%," according to one of the texts cited by the court.[8] The Hazelwood and Castaneda cases appear to attempt to reference the 5% and 1% criteria. The 5% level of statistical significance is approximately where the Supreme Court obtained its "two standard deviations" criteria, in place of the statistical criteria of 1.96 standard deviations. The 1% significance level occurs when the degree of disparity is 2.576 standard deviations, which is between two or three standard deviations.[9]

Yet neither of the most common significance levels in scientific testing, 1% or 5%, are "two or three standard deviations."

The "two standard deviations" criteria stated by the Supreme Court for tests of proportions results in a 4.55% chance of occurrence even when there is no discrimination. This is an uncommon, if ever used, significance level. While 2 may be close to 1.96 in some sense, it is not discussed in statistical texts as a criteria for statistical significance, even in those cited by the Supreme Court as justification for its criteria.

The "three standard deviations" criteria created by the Supreme Court represents a greater departure from statistical science and, because it has been followed by some courts, a greater damage to judicial decisions.[10] A degree of disparity of three standard deviations has only a 0.27% chance of occurring in the absence of discrimination. This is a much higher standard than 5% or 1%.

This means that if a court uses three standard deviations as the relevant criteria, it is requiring a degree of certainty that is far greater than is commonly used in statistical, scientific testing. In doing so, the court will underidentify statistical evidence of discrimination compared to common scientific standards.

How was the Supreme Court so wrong?

How did the Supreme Court miss the mark so badly in discussing these statistical standards, particularly when these statistical standards had been so well established in the statistical and scientific literature for decades?[11]

The degree of disparity in the Castaneda case was so great that the Supreme Court did not need to delineate the precise degree of disparity required for statistical significance. In Castaneda, the Supreme Court said the degree of disparity was as much as 29 standard deviations. As the Supreme Court stated, the "likelihood that such a substantial depature (sic) from the expected value would occur by chance is less than 1 in 10140."[12]

For the facts in Castaneda, whether the court standard for statistical significance was 1.96 standard deviations, as in the common scientific 5% criteria; two standard deviations, reflecting a virtually unused 4.55% criteria created by the Supreme Court; 2.576, reflecting the 1% highly statistically significant standard; or 3, which would be an uncommon and unused 0.27% Supreme Court criteria, was not relevant to the decision in the case because the degree of disparity was far beyond any of these levels.

Similarly in Hazelwood, the court did not define the degree of dispersion needed for prima facie evidence of discrimination. Instead, the court simply cited the Castaneda statement of "two or three standard deviations," mentioned that some of the case evidence resulted in a degree of dispersion of six standard deviations, and sent the case back to the lower court because the lower court "did not evaluate the factual record before it in a meaningful way."[13]

The error in the use of these statistical standards crept into wage discrimination cases through the invalid interpretation of the Supreme Court's statements in these cases as others applied them to new cases.

The results of Castaneda and Hazelwood on statistical evidence are far-reaching. These cases are often cited as the precedent for using statistical evidence in discrimination cases. At the same time, they have propagated standard benchmarks that are inconsistent with the same statistical science and texts that the Supreme Court cited in support of its discussion of the court's statistical benchmarks.

The impact of the court's description of the degree of disparity supporting prima facie evidence of discrimination becomes even more troubling when applied to other kinds of statistical analysis, such as regressions, commonly used in pay discrimination cases. Here, the measure of the degree of dispersion reflecting a given level of significance is not a constant, but rather varies depending on the number of people in the analysis, meaning that the use of the "two or three standard deviations" criteria would not only be errant, but the degree of the error would vary depending on the size of the group under analysis.

So, how should we interpret the Hazelwood/Castaneda standard for statistical evidence in discrimination testing?

The answer is simple. Courts and practitioners should ask the right question: "Is the difference between the plaintiff and benchmark groups significant at the 5% level?"

A court may choose some other significance level, such as 1%, depending on the situation in the case. However, courts and practitioners should simply use criteria based on the probabilistic levels of significance instead of the "two or three standard deviations" criteria

the Castaneda Supreme Court mentions in its attempt to roughly approximate the 5% and 1% levels of significance.

The accurate translation between "standard deviations" and statistical significance levels is known for each relevant statistical test, such as for proportions, means, regression coefficients and others. This statistical significance level is often called the p-value and is calculated automatically in statistical software for many statistical tests.

Statistics can be hard and confusing to many, and even the people on the Supreme Court can make a mistake. We should not hold the Supreme Court to an error where its intent was clear and a correction is in line with all the statistical science it cited as academic support for its statistical standard, even if it is a "depature" (sic) from its precise wording.

Daniel S. Levy, PhD, is national managing director at Advanced Analytical Consulting Group Inc. Dr. Levy is also the CEO of EquiCalc Inc.

[1] EEOC, WHD, Memorandum of Understanding Between the U.S. Department of Labor, Wage and Hour Division and the U.S. Equal Employment Opportunity Commission, Sept. 13, 2023.

[2] OFCCP, Directive (DIR) 2022-01 Revision 1. August 2023.

[3] Castaneda v. Partida, 430 U.S. 489 (1977).

[4] Hazelwood School District v. United States (1977), 433 U.S. 299.

[5] 433 U.S. n. 14 citing 430 U.S. 497 n. 17.

[6] Finkelstein, The Application of Statistical Decision Theory to the Jury Discrimination Cases, 80 Harv. L. Rev. 338, 353-356 (1966). See generally P. Hoel, Introduction to Mathematical Statistics 58-61, 79-86 (4th ed. 1971); F. Mosteller, R. Rourke, & G. Thomas, Probability with Statistical Applications 130-146, 270-291 (2d ed. 1970).

[7] The Hazelwood opinion noted a disagreement about the correct proportion of African American teachers in the relevant population. We use the 5.7% figure mentioned by the court for ease of exposition.

[8] F. Mosteller, R. Rourke, & G. Thomas, Probability with Statistical Applications, p. 311.

[9] This is based on the normal approximation to the binomial, which may be valid in "large samples" as referenced by the Supreme Court.

[10] See for example OFCCP, DOL v. Analogic Corp., Recommended Decision and Order, P. 40, where the administrative law judge found that while a "wage disparity at 2.84 standard deviations is statistically significant, it does not reflect a 'gross' statistical disparity such that the statistics alone are enough to satisfy OFCCP's burden." The chance of a degree of

dispersion this large in the population studied in the absence of underlying disparity was less than 5 out of 1,000, ten times more stringent than the common 5% significance level and twice the most stringent level used scientific research, the 1% significance level. The dollar value of the difference was over 80 cents per hour, controlling for explanatory factors, for women with an hourly wage of approximately $20.

[11] See for example, the academic authorities cited by the Supreme Court in n. 6 above and C. J. Clopper, E. S. Pearson, "The Use of Confidence Or Fiducial Limits Illustrated in the Case of the Binomial," Biometrika, Volume 26, Issue 4, December 1934, Pages 404-413, https://doi.org/10.1093/biomet/26.4.404. Clopper, Pearson reflects the development of "exact" tests for binomials. In some contexts it is important to note that the confidence intervals for a binomial test are increasingly non-symmetric as the resulting value moves away from 50% toward either 0% or 100%. This means that using any "standard deviation"-based test that resulted in symmetric confidence intervals around the observed binomial proportion would be inaccurate except at 50%. "Exact" tests documented since the 1930s account for this non-symmetry. All tests of binomial proportions based on any number of standard deviations do not account for it.

[12] Castaneda v. Partida, 430 U.S. 482 (1977), n. 17. Based on alternative figures, the Castaneda court noted the degree of dispersion might be 12 standard deviations. Again, it is well beyond "two or three standard deviations" that the Castaneda court references for statistical significance.

[13] Hazelwood School District v. United States, 433 U.S. 299 (1977).