1-1-2017

Pernicious P-Values: Statistical Proof of Not Very Much

Kingsley R. Browne
Wayne State University

PERNICIOUS *P*-VALUES: STATISTICAL PROOF OF NOT VERY MUCH

Kingsley R. Browne[*]

I. INTRODUCTION

Statistical evidence plays a large role in employment discrimination cases. It is almost essential in both pattern-and-practice cases[1] and disparate-impact cases,[2] and it is sometimes used in support of individual disparate-

[*] Professor, Wayne State University Law School. E-mail: kingsley.browne@wayne.edu. My thanks to April Bleske-Rechek, Tony Dillof, John Dolan, John Rothchild, and participants in a brownbag discussion at Wayne State University Law School, for comments on a prior draft of this paper.

[1] *See, e.g.*, Int'l Bhd. of Teamsters v. United States, 431 U.S. 324 (1977).

[2] *See, e.g.*, Griggs v. Duke Power Co., 401 U.S. 424 (1971).

treatment cases.[3] Hundreds of reported employment discrimination cases involve the use of statistical evidence,[4] and in many cases the evidence is treated as largely dispositive, at least for purposes of establishing a prima facie case or certifying a class, and often for liability itself. Numerous law-review articles have been devoted to the topic, and most law school casebooks in the field cover issues of statistical proof.

The typical statistical case relies on what is known as "null hypothesis significance testing" (hereinafter "NHST"). When the demographic group of interest is not proportionally represented in some component of an employer's workforce, the question is whether the under-representation is so great that it should be concluded that the disparity is not just a result of chance (like getting only four heads in ten coin tosses, rather than the statistically "expected" five). The inference that the plaintiff would like drawn is that the observed disparity is so great that chance is not a likely explanation and that an impermissible criterion (*e.g.*, race or sex) is instead to blame. Using NHST, the plaintiff's statistical expert will test the "null hypothesis" that hiring (or promotion, etc.) is random with respect to, say, race—in that blacks and whites in the relevant groups have an equal chance of being selected—and reject that hypothesis if the disparity is "statistically significant," usually meaning that the results are such that would be found only among less than 5% of employers hiring at random with respect to race. Rejection of the null hypothesis leads to acceptance of the alternative hypothesis, of which more later.

One might expect that even if the nuances of statistical analysis are sometimes misunderstood by legal professionals, at least there would be broad consensus on what might be called "the basics." Unfortunately, to the extent that there is broad consensus on the basics, the consensus is wrong. In fact, courts routinely make fundamental errors in their interpretation of statistical evidence.[5]

There are a number of problems with NHST, but one basic error that courts regularly make is what has been called the "transposition fallacy."[6] In the context of a discrimination lawsuit, one commits the transposition fallacy by equating the probability of an observed statistical disparity given random selection with the probability of random selection given the observed

---

[3] *See, e.g.*, Baylie v. Federal Reserve Bank of Chicago, 476 F.3d 522 (7th Cir. 2007) (holding that statistical evidence is relevant, but only marginally so, in individual disparate-treatment cases); Obrey v. Johnson, 400 F.3d 691 (9th Cir. 2005) (holding that such evidence is relevant and therefore admissible).

[4] A Westlaw search of the "All Federal" database on September 6, 2017, found 500 cases satisfying the query ("Title VII" and "statistically significant") and 3,530 satisfying the query ("Title VII" and "statistical evidence").

[5] *See generally* Kingsley R. Browne, *Statistical Proof of Discrimination: Beyond "Damned Lies,"* 68 WASH. L. REV. 477 (1993) [hereinafter Browne, *Beyond Damned Lies*].

[6] *See* Kingsley R. Browne, *The Strangely Persistent "Transposition Fallacy": Why "Statistically Significant" Evidence of Discrimination May Not Be Significant*, 14 LAB. LAW. 437 (1998); HANS ZEISEL & DAVID H. KAYE, PROVE IT WITH FIGURES: EMPIRICAL METHODS IN LAW AND LITIGATION 82 (1997).

disparity.[7] What might seem to be a mere technical error—confusion between two things that sound very much alike—has substantial real-world effects, because it causes courts to mistakenly conclude that they can quantify the likelihood that the employer's selection process was random with respect to the trait at issue. In many cases, the courts' conclusions are understandable, because they are led into their error, whether wittingly or unwittingly, by the parties' experts.

Even if the transposition fallacy could be avoided—which seems unlikely given its persistence in the face of decades of scholarly criticism—a more comprehensive problem with much statistical proof is its reliance on NHST in the first place. As will be discussed below, there is a widespread and growing dissatisfaction with the use of NHST, partly because of the ease with which it leads to the transposition fallacy, but also because of other weaknesses, such as its tendency to lead researchers to conflate statistical and practical significance and to erroneously conclude that rejection of the null hypothesis necessarily demonstrates the truth of some specified alternative hypothesis.

This article will proceed in two parts. First, I will discuss the ubiquitous transposition fallacy, showing how it misleads courts and why it matters. Second, I will describe some of the other problems with NHST, which make it an inappropriate basis for determining liability.

II. THE TRANSPOSITION FALLACY

A. What is the Transposition Fallacy?

Formally, the transposition fallacy, which goes by several names,[8] entails equating "the probability of A given B" with "the probability of B given A." Consider two statements: (1) "there is a 99.9% chance that the animal will have two arms and two legs given that it is a chimpanzee (the uncertainty deriving from the possibility of a birth defect or amputation); and (2) "there is a 99.9% chance that the animal is a chimpanzee given that it has two arms and two legs." The latter proposition obviously does not follow

---

[7] ZEISEL & KAYE, *supra* note 6, at 82.

[8] *See, e.g.*, NATIONAL RESEARCH COUNCIL, THE EVALUATION OF FORENSIC DNA EVIDENCE 133 (1996) (referring to the fallacy as "the fallacy of the transposed conditional"); Browne, *Beyond Damned Lies*, *supra* note 5, at 484–503 (referring to the fallacy as "the statistical fallacy"); David H. Kaye & Jonathan J. Koehler, *Can Jurors Understand Probabilistic Evidence?*, J. ROYAL STAT. SOC'Y, Series A 75, 77 (1991) (referring to the fallacy as the "inversion fallacy"); William C. Thompson & Edward L. Schumann, *Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy*, 11 LAW & HUM. BEHAV. 167, 181 (1987) (referring to the fallacy as the "prosecutor's fallacy"); GEOFF CUMMING, UNDERSTANDING THE NEW STATISTICS: EFFECT SIZES, CONFIDENCE INTERVALS, AND META-ANALYSIS 27, 27 (referring to the fallacy as the "inverse probability fallacy"); Thomas Sellke, M. J. Bayarri, & James O. Berger, *Calibration of* p *Values for Testing Precise Null Hypotheses*, 55 AM. STATISTICIAN 62, 62 (2001) (referring to the fallacy as the "*p* value fallacy") [hereinafter Sellke et al.]; Jacob Cohen, *The Earth is Round (p < .05)*, 49 AM. PSYCHOL. 997, 999 (1994) (referring to the fallacy as the "inverse probability error").

from the former, and no one would for a minute argue that it did. Yet arguments with exactly that form abound in legal cases.

1. The Prosecutor's Fallacy

Before examining the transposition fallacy in the context of employment discrimination cases—where the fallacy is both ubiquitous and unrecognized by the courts—it might be useful to describe an area of the law where the fallacy is much more widely recognized: the use of DNA profiles in criminal prosecutions. Indeed, the fallacy has its own name in such cases, where it goes by the name of the "prosecutor's fallacy" because of the unfair advantage that it gives prosecutors in DNA cases.

Imagine that forensic investigators at the scene of a murder find a blood sample that they believe came from the perpetrator. A DNA profile from that blood is compared to a DNA database and yields a "cold hit" on Innocent Irving, who unfortunately does not have an alibi for the time when the crime was committed. The likelihood that a random person who was not the source of the crime-scene DNA sample would have matching DNA is 1 in 10,000. Bingo!

The prosecutor would like to argue to the jury that, given the infrequency of the DNA profile, there is only a 1 in 10,000 chance that the finding was coincidental and a 9,999 in 10,000 chance that Irving was the actual source of the crime-scene sample. In other words, the prosecutor would equate the probability that a sample of DNA from the crime scene would match a randomly selected person *given that he is not the source of the crime-scene DNA* (1 in 10,000 in our case), with the probability that he is not the source of the crime-scene DNA *given that there is a match.*[9] This fallacious argument is a powerful one for the prosecutor, indicating to the jury that the probability that Irving is the source of the DNA is .9999 (and, of course, the prosecutor would like the jury to make the next inference, which is that if he was the source of the crime-scene DNA, he is necessarily the murderer).

The error in the prosecutor's argument has seemed relatively obvious to courts, many of which have described the prosecutor's fallacy and pointed out its error,[10] despite the fact that the prosecutor's fallacy embodies the same

---

[9] *See* DAVID H. KAYE & GEORGE SENSABAUGH, *Reference Guide on DNA Identification Evidence*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE, 168 n.88 (3d ed. 2011); Thompson & Schumann, *supra* note 8, 170–71.

[10] The United States Supreme Court has described the prosecutor's fallacy as follows:

> The prosecutor's fallacy is the assumption that the random match probability is the same as the probability that the defendant was not the source of the DNA sample. . . . [I]f a juror is told the probability a member of the general population would share the same DNA is 1 in 10,000 (random match probability), and he takes that to mean there is only a 1 in 10,000 chance that someone other than the defendant is the source of the DNA found at the crime scene (source probability), then he has succumbed to the prosecutor's fallacy. It is further error to equate source probability with probability of guilt, unless there is no explanation other than guilt for a person to be

flawed reasoning that leads courts to conclude in discrimination cases that the *p*-value provides the probability of random selection.[11] For some reason, what seems at least somewhat widely appreciated in the criminal context[12] seems not to penetrate other contexts.

2. The Transposition Fallacy in Discrimination Cases

The equivalent of the prosecutor's fallacy is a staple of discrimination cases. A concrete example of the fallacy in the discrimination context will illustrate. Suppose that a given employer has 500 employees and that blacks constitute 20 percent of the qualified labor force in the relevant market. We would expect statistically that 20% (100 employees) of the employer's workforce would be black. When we say that we would "expect" the employer to have 20 percent black employees, we do not mean that we would actually believe that every employer would have exactly that number, any more than we would expect that every time a person flipped a coin 100 times the result would be exactly 50 heads and 50 tails. Instead, what we mean is that out of a very large number of non-discriminating employers, we would expect the *average* representation of black employees to be 20%, just as we would expect that, over the course of many instances of fair coins being flipped 100 times, the average result would be 50:50. Very few employers of 500 people (less than 5 percent) would actually be predicted to have exactly 100 black employees, just as few people (less than 8 percent) who toss a coin 100 times would be predicted to get exactly 50 heads.[13]

> the source of crime-scene DNA. This faulty reasoning may result in an erroneous statement that, based on a random match probability of 1 in 10,000, there is a .01% chance the defendant is innocent or a 99.99% chance the defendant is guilty.

McDaniel v. Brown, 558 U.S. 120, 128 (2010).

In *McDaniel*, the prosecution expert had testified that in light of the 1 in 3,000,000 random-match probability, it was "not inaccurate" to say that the likelihood that the sample had not come from the defendant was .000033. *Id.* Moreover, the prosecutor had argued in his closing that the jury could be "99.999967 sure" that the DNA came from the defendant. *Id.* at 128–29.

[11] The same DNA statistics can be misused by the defendant as well, of course. The defendant might argue, for example, that in a city of 1,000,000 people, a hundred would be expected to match the sample. Based *solely* on the DNA evidence, then, one might well argue that the chance the sample came from the defendant–rather than being 9,999 out of 10,000—is actually 1 out of 100. That would be true, of course, only if the *only* evidence in the case is the DNA frequency evidence and all persons having matching DNA could potentially be the culprit (which, of course, would presumably not be true of small children). To make the one-in-a-hundred argument in the face of other evidence tying the defendant to the crime is to engage in the "defense attorney's fallacy." *See* Thompson & Schumann, *supra* note 8, at 171.

[12] *See, e.g.*, McDaniel v. Brown, 558 U.S. at 128; U.S. v. Chischilly, 30 F.3d 1144, 1153 (9th Cir. 1994); U.S. v. Shea, 957 F. Supp. 331, 332 (D. N.H. 1997); U.S. v. Pritchard, 993 F. Supp. 2d 1203, 1209 (C.D. Cal. 2014); State v. Jackson, 221 P.3d 1213, 1222 (Mont. Sup. Ct. 2009); State v. Ragland, 739 S.E.2d 616, 623 (N.C. App. 2013) (agreeing with defendant's contention that the state's expert committed the "prosecutor's fallacy" but affirming conviction because other overwhelming evidence indicated that admission was not "plain error"); Brown v. Farwell, 525 F.3d 787, 795–796 (9th Cir. 2008), *rev'd*, 558 U.S. 120 (2010) (affirming grant of habeas corpus on the basis of expert testimony equating random-match probability with source probability).

[13] These probabilities can be calculated using an online binomial calculator. *Binomial Calculator: Online Statistical Table*, STAT TREK, http://stattrek.com/online-calculator/binomial.aspx (last visited May 1 2017).

Suppose, however, that our employer has 82 black employees instead of the "expected" 100. Clearly, it has fewer than expected, but is it sufficiently fewer that we ought to be suspicious that some nefarious cause—or even just a systematic cause that may or may not be nefarious—is responsible? The statistical analysis used in employment discrimination cases is intended to answer that question, although it answers that question substantially less well than is generally assumed.

In order to test this question, the statistician will use null hypothesis significance testing. The "null hypothesis" that is tested is that there is no difference in the likelihood of selection of blacks versus whites;[14] that is, the composition of the employer's workforce is a product of forces that are purely random with respect to race, such that any differences that do exist are a product of sampling error (that is, the luck of the particular draw). The alternative hypothesis, which would be accepted if the null hypothesis is rejected, is that there is in fact a racial difference in the probability of selection. That is, there is something systematic leading to the racial disparity (and, of course, the plaintiff would argue that the "something" is racial discrimination, though rejection of the null hypothesis does not lead ineluctably—or even necessarily very strongly—to that conclusion[15]).

---

[14] In practice, the null hypothesis is generally a hypothesis of no difference, but it does not have to be so. The null hypothesis is any hypothesis to be "nullified." Gerd Gigerenzer, *Mindless Statistics*, 33 J. SOCIO-ECON., 587, 589 (2004). Some statisticians refer to a null hypothesis of no difference as the "nil hypothesis." *See, e.g.*, REX B. KLINE, BEYOND SIGNIFICANCE TESTING: STATISTICS REFORM IN THE BEHAVIORAL SCIENCES 69, 69 (2013).

[15] *See infra* Sections III.A, III.D. It is also important to note that contrary to occasional statements to the contrary, failing to reject the null hypothesis does not entail an "acceptance" of it. That is, one can never *prove* the null hypothesis. *See* CUMMING, *supra* note 8 at 29 ("We must therefore be careful not to take statistical nonsignificance (the null is not rejected) as evidence of a zero effect (the null is true)."). Cumming cautions against the "slippery slope of non-significance," whereby "[a]n effect is found to be statistically nonsignificant then later discussed as if that showed it to be zero"). *Id.* For examples of this error, see STEPHANIE R. THOMAS, STATISTICAL ANALYSIS OF ADVERSE IMPACT 45 (2011) ("If there is no statistically significant difference between the actual and expected outcomes experienced by protected and non-protected individuals, the null hypothesis cannot be rejected. *One would infer that the null hypothesis is true, and conclude that there is no relationship between protected status and the outcome of the challenged employment practice or policy.*" (emphasis added)); Report of Helen Reynolds, Ph.D., Ellis v. Crawford, No. 3:03-cv-02416, 2005 U.S. Dist. LEXIS 3457 (N.D. Tex. Mar. 3, 2005) ("The significance level (called "p") of 5 percent is often used as the cut-off point. *If p is greater than 0.05, then one can say that the two amounts are not really different.* That is, if p is 0.05 or greater, there is, at minimum, a 5 percent (one in twenty) chance that, all else equal, the observed difference is due to chance and chance alone. If that is the case, *we would accept the null hypothesis that there was no statistical difference in the two numbers*" (emphasis added)); MICHAEL J. ZIMMER, CHARLES A. SULLIVAN, & REBECCA HANNER WHITE, CASES AND MATERIALS ON EMPLOYMENT DISCRIMINATION 134 (8th ed., 2013) [hereinafter ZIMMER ET AL., 8th ed.] (noting that "[t]he employer would prefer to *confirm the null hypothesis–that is, to show that any difference is due to chance*" (emphasis added)); *id.* at 133 ("To use probability theory to prove discrimination, a statistician would construct an assumption, called the null hypothesis, which would then be tested and *either accepted or rejected.*" (emphasis added)).

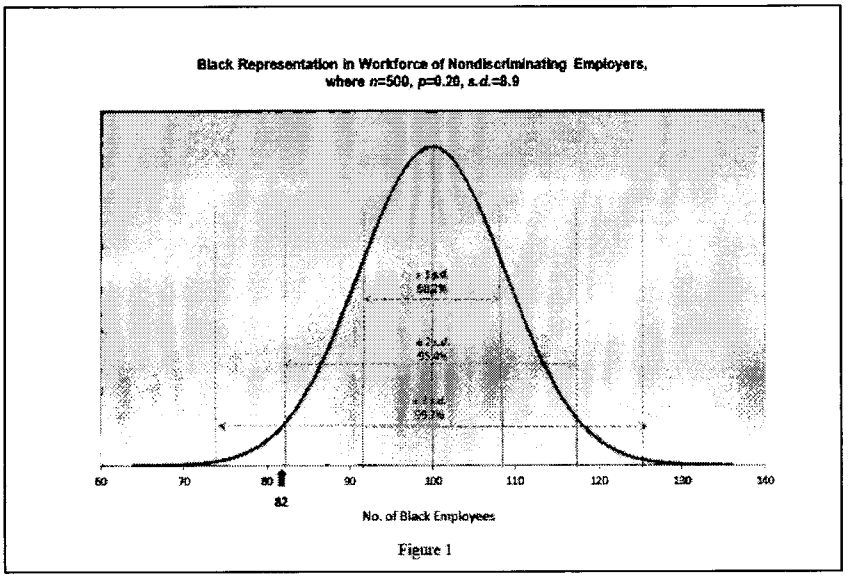

Figure 1

The distribution of the demographic profile of employers' workforces can be represented graphically by the "normal" or "bell-shaped" curve. Figure 1 is a representation of a hypothetical distribution of employers of 500 employees each, who select their employees randomly with respect to race from a very large pool of applicants of whom 20 percent are black. The results cluster around a mean of 100 black employees, and as one moves farther away from the center of the distribution—so that the black percentage becomes increasingly far from the "expected" 20%—the number of employers decreases. Because these data are normally distributed, one can calculate the percentage under any part of the curve if the standard deviation of the data set is known. If we compute the standard deviation in the case of our hypothetical distribution of non-discriminating employers, we can then estimate the likelihood that a randomly chosen employer will have a particular racial composition, *assuming that the null hypothesis is true*.

The standard deviation is readily calculated. The formula for the standard deviation (using the normal approximation of the binomial distribution) is

$$s.d. = \sqrt{np(1-p)}$$

where n is the number of trials, and p is the probability of obtaining the result in a given trial. In this case, $n = 500$ and $p = 0.2$, so that the standard deviation is:

$$s.d. = \sqrt{500 * 0.2(1-0.2)} = 8.9$$

The next step is to calculate the *z-score*, which is the number of standard

deviations by which our employer's workforce deviates from the mean:[16]

$$z = \frac{observed - expected}{s.d.}$$

or

$$z = \frac{82 - 100}{8.9} = -2.02$$

The final step is to determine the probability associated with a z-score of ± 2.02. A standard probability table or online calculator reveals that the probability associated with a z-score of ± 2.02 is .043, meaning that there is less than a five percent chance that any given non-discriminating employer chosen at random would have this great (or greater) a deviation from the mean if the null hypothesis is true.[17] This probability is commonly referred to as the "p-value."[18] The results would thus be "statistically significant" using the conventional five-percent significance level, although they would not be if a more stringent significance level of one-percent were used.[19]

---

[16] In any normal curve, 68% of the distribution will fall within one standard deviation of the mean, 95% will fall within two standard deviations, and 99.7% will fall within three. *See* Figure 1. Translated to our hypothetical distribution, that means that about 68% of randomly selecting employers would be expected to have between 91 and 109 black employees, 95% would have between 82 and 118 black employees, and 99.7% would have between 73 and 127 employees.

[17] *See, e.g.*, David M. Lane, *Normal Distribution*, ONLINE STATISTICS EDUCATION: AN INTERACTIVE MULTIMEDIA COURSES STUDY, http://onlinestatbook.com/2/calculators/normal_ dist.html (last visited May 1, 2017). This discussion assumes that a "two-tailed test" is being used. Under a two-tailed test, the null hypothesis is that there is no racial difference (*i.e.*, blacks have the same probability of being hired as whites do), and the alternative hypothesis is that the probability of being hired is different, although the difference could run in either direction. Under a one-tailed test, the alternative hypothesis is that the difference runs in a particular direction (*e.g.*, blacks have a lower probability of being hired than whites). Another way of saying this is that in a two-tailed test, the statistically significant results (5% in this case) are distributed between the two tails of the distribution, with 2.5% in each tail. In a one-tailed test, the significant results are confined to one tail. As a result, using a two-tailed test, a z-score of ± 1.96 is necessary to declare a result statistically significant at the 5% level (hence the two-standard-deviation disparity that many courts look for), while using a one-tailed test, a z-score of only 1.65 is required, but it must be of the previously designated sign. There is an ongoing dispute in the literature and the cases about whether a one-tailed or two-tailed test is appropriate in this context. DAVID H. KAYE & DAVID A. FREEDMAN, *Reference Guide on Statistics*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 211, 255–56 (3d ed. 2011).

[18] *See* CUMMING, *supra* note 8, at 22. Because half the employers in the hypothetical distribution with a disparity this large would have an "excess" of blacks, the probability that an employer in this distribution would have 82 or fewer blacks is half that amount, or .0215).

[19] Although the focus of this article is on binomial statistics, the same issues arise with multiple-regression analyses. Multiple-regression analysis is often used in cases alleging discrimination in compensation. It attempts to control for legitimate factors that are thought to contribute to salary and determine whether, after having done so, there is still a difference between the groups, which the model then attributes to the "dummy variable" of sex, race, etc. *See* KAYE & FREEDMAN, *supra* note 17, at 279–81. The discrimination inquiry then focuses on whether the coefficient of the dummy variable is statistically significant, relying on the same p-values as in the binomial analysis. The quality of the regression analysis rests heavily on selection of the appropriate variables. Omitting a variable that differs between the groups of interest can lead to spurious findings of discrimination. For example, failing to include the number of hours worked in a regression comparing annual earnings of men and women would result in any sex

When we see an employer with a workforce demographic like our hypothetical one, what can we say about the likelihood that something systematic (whether discrimination or something else, such as differential qualifications or interest) is responsible for the racial imbalance? The typical answer provided in the case law is that a *p*-value of less than .05 means there is less than a five-percent chance that the disparity was caused by chance (with a corresponding probability of greater than 95 percent that it had a nonrandom cause).[20] However, equating the *p*-value with the probability of random selection (in other words, equating the *p*-value with the probability that the null hypothesis is true) embodies the transposition fallacy.[21] So, the correct answer to the question is that by itself the *p*-value tells us "very little" (although perhaps not nothing[22]) about that probability.[23]

Why is it an error to equate the *p*-value with the probability of random selection—that is, the probability that the null hypothesis is true? The answer gets back to what the distribution is that is described in Figure 1. It is a distribution of employers whose workforces were assembled at random with respect to race. That is, *all* of the employers represented in that distribution had an equal probability of selecting whites and blacks; the distribution is the distribution *assumed* by the null hypothesis. While the probability figure does represent the probability of an individual *non-discriminating* employer selected at random obtaining such a distribution if the null hypothesis is true, it does not reveal the likelihood that the null hypothesis is true. Stated otherwise, "*the P value is calculated on the assumption that the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false.*"[24]

B. An Illustration of the Error of the Transposition Fallacy

The lack of equivalence between the *p*-value and the probability of discriminatory (or even nonrandom) selection can be easily seen by consideration of a few hypothetical situations, with the only difference between them being the known frequency of discrimination among employers. For ease of discussion, we will assume that the only possible systematic cause of racial disparities is

---

difference in earnings that is caused by differences in hours worked to be attributed to sex. *See* DANIEL L. RUBINFELD, *Reference Guide on Multiple Regression*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 303, 313–16 (3d ed. 2011).

[20] *See infra* notes 30–33 and accompanying text.

[21] *See* Regina Nuzzo *Scientific Method: Statistical Errors*, 506 NATURE 150, 151 (2014), www.nature.com/news/scientific-method-statistical-errors-1.14700 (last visited May 1, 2017) ("Most scientists would look at" results with a "*P* value of .01 and say that there was just a 1% chance of [the] result being a false alarm. But they would be wrong. The *P* value cannot say this: all it can do is summarize the data assuming a specific null hypothesis.").

[22] *See infra* Section II.E.

[23] CUMMING, *supra* note 8, at 28 (noting that although *p* is not the probability that the null is true, "[s]urprising results may reasonably lead you to doubt the null").

[24] Steven N. Goodman, *Toward Evidence-Based Medical Statistics. 1: The* P *Value Fallacy*, 130 ANN. INTERN. MED. 995, 998 (1999).

discrimination (although that is, in most cases, a highly implausible assumption),[25] and also that all discriminating employers would have a racial disparity that is statistically significant.

Consider the following scenario: there are 10,000 employers in a region in which there is a fair amount of racial diversity. Obviously, the workforces of few, if any, employers are going to match the "expected" racial composition exactly, and some, even non-discriminating ones, would be expected to exhibit major deviations from the statistically expected profile. In fact, by definition, five percent (or 500) of them would be expected to have disparities that are statistically significant at the five-percent level if they were selecting at random. So, what is the probability that a given employer with a statistically significant disparity obtained it by chance? If you adhere to the transposition fallacy, the answer is always five percent. If you do not, the answer is, "it depends on the frequency of discrimination in the population," as the following five scenarios reveal.[26]

> 1. **Base Rate = 0.** Suppose you know that no employers discriminate (that is, the base rate of discrimination in this population is zero). Then, the answer is—by definition—that *none* of the employers with statistically significant disparities discriminates, notwithstanding the *p*-value of .05; all will have obtained the result by chance. (See Figure 2.)

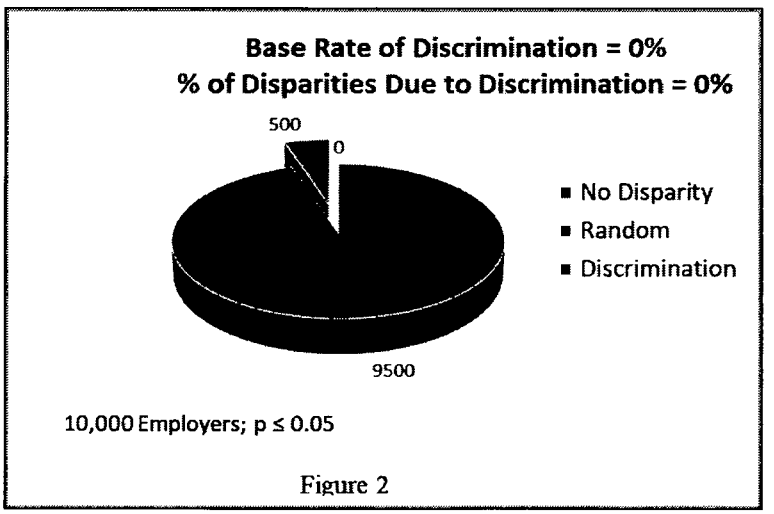

Figure 2

[25] *See infra* Sections III.A, III.D.

[26] *See infra* Figure 2 in Section II.B. It is important to note that when we say "frequency of discrimination" here we mean the frequency of employers who engage in a pattern or practice of discrimination, not "isolated instances" of discrimination. *See* Teamsters v. United States, 431 U.S. 324, 336 (1977) (noting that government in a pattern or practice case must show that "racial discrimination was the company's standard operating procedure - the regular rather than the unusual practice" and that it must "prove more than the mere occurrence of isolated or . . . sporadic discriminatory acts").

2. **Base Rate = .01.** What if, instead, you know that one percent of employers discriminate? Then, there will be 100 disparities due to discrimination (1% of 10,000) and 495 due to chance (5% of the remaining 9900). So, in this example, 16.9% (100/595) of the observed disparities will be caused by discrimination. (See Figure 3.)

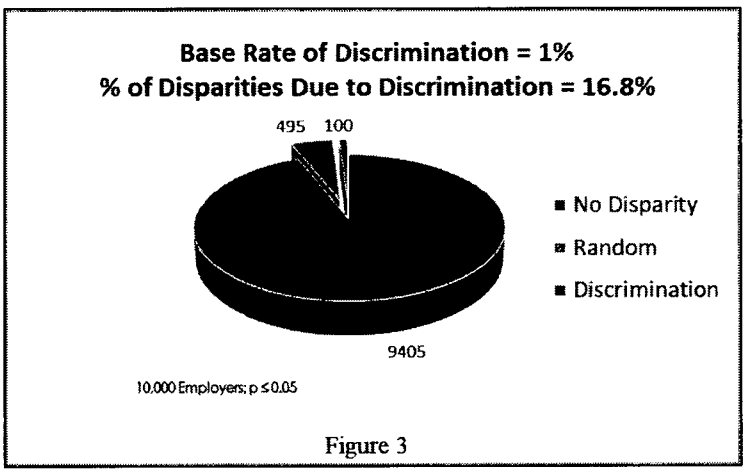

Figure 3

3. **Base Rate = .05.** What if, instead, five percent of employers discriminate? Then, 500 of them will have obtained their disparities through discrimination, and 475 (5% of the remaining 9500) will have obtained their disparities by chance. Now, the likelihood that an employer with a statistically significant discrepancy obtained it through discrimination is 500/975 or 51.3%—more likely than not (barely) but a far cry from the 95% figure that the transposition fallacy would suggest. (See Figure 4.)

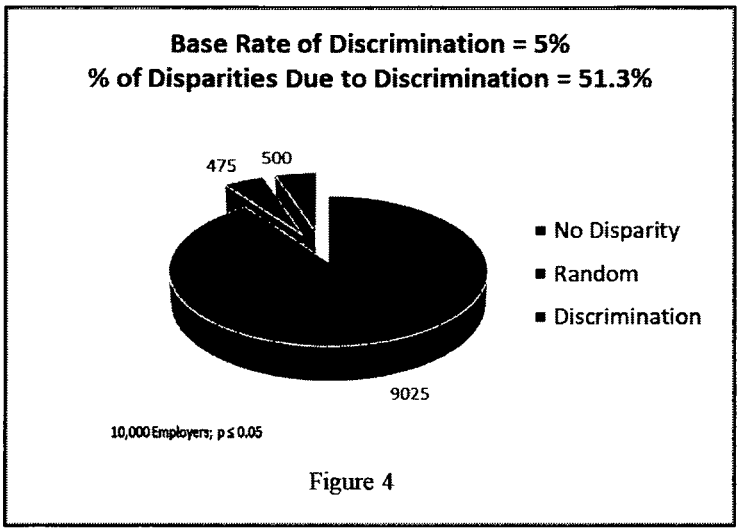

Figure 4

4. **Base Rate = .10.** Suppose ten percent of employers discriminate. Then, 1,000 (10%) will have disparities due to discrimination, and 5% of the remaining 9,000 (450) will have disparities due to chance, such that 69% of employers with disparities will have obtained them through discrimination. (See Figure 5.)

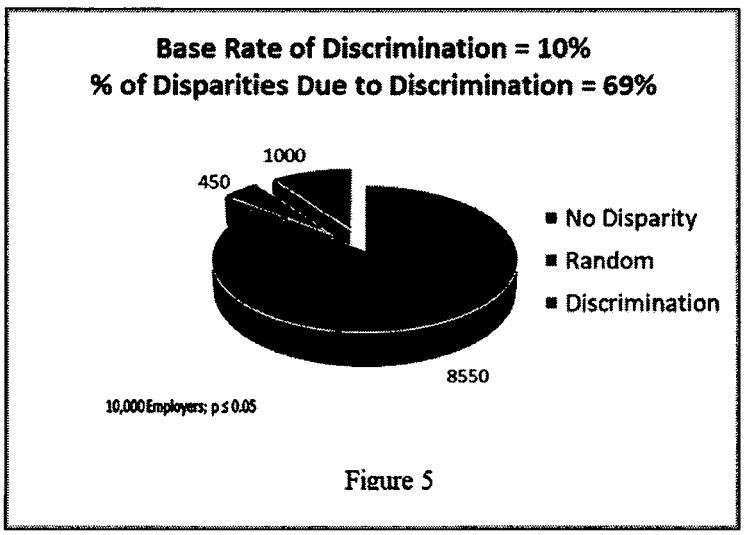

Figure 5

5. **Base Rate = .50.** Finally, suppose that half of all employers engage in systematic racial discrimination. Now,

5,000 employers would have statistically significant disparities from discrimination, and 250 employers (5% of the remaining 5000) would have disparities caused by chance. Only now—with an assumption that half of all employers systematically discriminate—does the percentage of disparities due to nonrandom factors (95.3%) approximate the percentage *assumed* under the transposition fallacy, and that is just by happenstance.[27] (See Figure 6.)

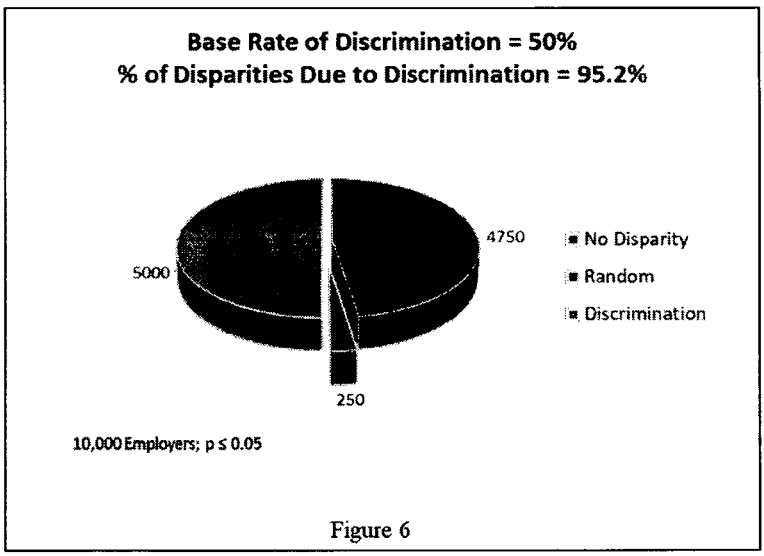

Figure 6

In all five scenarios, the *p*-value supplied by NHST is the same—the conventional .05 level. Yet, as we can see, the probability supplied by the *p*-value has virtually no relationship to the probability that a given employer with a statistically significant disparity obtained that disparity by chance.[28] Moreover, in real life, of course, there will be many additional disparities that are due to systematic, yet nondiscriminatory, causes, such as differential

[27] These figures assume that any discrimination could run in either direction; *i.e.*, either favoring whites or favoring blacks, which is the assumption of the two-tailed test. Only one-half of the disparities described above would disadvantage blacks if a two-tailed test is used. If a one-tailed test is used, then all of the disparities would disadvantage blacks.

[28] There is a more formal way to calculate the probabilities incorporating the base-rate of discrimination into the analysis, and that is to use Bayesian analysis, which does not rest on the transposition fallacy. *See generally* Jason R. Bent, *Hidden Priors: Toward a Unifying Theory of Systemic Disparate Treatment Law*, 91 DENVER U. L. REV. 807 (2014); *see generally* Deborah M. Weiss, *The Impossibility of Agnostic Discrimination Law*, 2011 UTAH L. REV. 1677 (2011). There are a number of problems with that approach, however, including the impossibility of ascertaining the base rate, as well as the inappropriateness of basing an employer's liability on estimates of the rate of wrongdoing of other employers. *See* Kingsley R. Browne, *No Bayesian Solution to the Transposition Fallacy* (manuscript on file with the author) (arguing that use of base rates of discrimination runs afoul of Rule 404 of the Federal Rules of Evidence).

qualifications or interest.[29]

C. *The Pervasiveness of the Transposition Fallacy*

That courts occasionally make errors in their interpretation of statistics is not remarkable or even terribly concerning. What is both remarkable and concerning is the fact that, whenever a court explains with any precision what a *p*-value means, it is virtually *always* wrong, as the many cases cited herein attest. For example, the Ninth Circuit recently explained that "a threshold of two standard deviations corresponds roughly to a 95% confidence level or a .05 level of significance, i.e., there is only a 5% probability that the result is due to chance."[30] Similarly, according to the Seventh Circuit, "two standard deviations is normally enough to show that it is extremely unlikely (that is, there is less than a 5% probability) that the disparity is due to chance, giving rise to a reasonable inference that the hiring was not race-neutral"[31] The Third Circuit has declared that "[p]robability

---

[29] One response to the transposition fallacy argument in discrimination cases is "why [should] the legal system . . . conclude that base-rate discrimination is especially rare"? ZIMMER ET AL, 8th ed., *supra* note 15, at 135. That argument misses the point. The point is not that discrimination is more or less rare than a particular person might assume but rather that the presumed meaning of the *p*-value is *absolutely incorrect*. A *p*-value of .05 equates to the 95% chance of discrimination that the transposition fallacy suggests only if the prevalence of systemic discrimination just happens to be about 50%. Yet, the 5% significance level was not adopted based on any particular assumption about base rates of discrimination but rather because it is a traditional significance level used in the social sciences that is erroneously believed to supply the probability of random selection.

[30] Brown v. Nucor Corp., 576 F. 3d 149, 156 n. 9 (4th Cir. 2009); *see also* Apsley v. Boeing Co., 691 F.3d 1184, 1197 (10th Cir. 2012) ("The odds of such a difference occurring [by chance] can be referred to as the probability, or *p*-value."); King v. Burlington N. Santa Fe Ry. Co., 762 N.W.2d 24, 37 (Neb. Sup. Ct. 2009) ("A significance level of .05 presents a 5-percent probability that researchers observed an association because of chance variations."); Dukes v. Wal-Mart Stores, Inc., 222 F.R.D. 137, 156 n.23 (N.D. Cal. 2004) ("The standard deviation is a number that quantifies the probability that chance is responsible for any difference between an expected outcome and the observed outcome . . . ."); Wynn v. Nat'l Broad. Co., Inc., 234 F. Supp. 2d 1067, 1124 n.47 (C.D. Cal. 2002) ("Generally, a statistical study is viewed as statistically significant if the probability that the result is attributable to chance alone is less than 1 out of 20." (citation omitted)).

[31] Adams v. Ameritech Servs., Inc., 231 F.3d 414, 424 (7th Cir. 2000); *see id.* at 427 ("The difference was statistically significant in six of those 12 categories, in the sense that there was a small (less than 5%) probability that the difference was due to chance."). *See also* Meditz v. City of Newark, 658 F.3d 364, 372 n.13 (3d Cir. 2011) ("The standard deviation calculation measures how likely it is that a deviant result occurred by chance." (quoting Ramseur v. Beyer, 983 F.2d 1215, 1232 n.17 (3d Cir. 1992))); Smith v. Xerox, 196 F.3d 358, 366 (2d Cir. 1999) ("If an obtained result varies from the expected result by two standard deviations, there is only about a 5% probability that the variance is due to chance."); Peightal v. Metropolitan Dade County, 26 F.3d 1545, 1556 n.16 (11th Cir. 1994) ("[S]ocial scientists consider a finding of two standard deviations significant, meaning there is about one chance in 20 that the explanation for the deviation could be random and the deviation must be accounted for by some factor other than chance." (quoting Waisome v. Port Authority, 948 F.2d 1370, 1376 (2d Cir.1991))); Jones v. Pepsi-Cola Metro. Bottling Co., 871 F. Supp. 305, 309 n.10 (E.D. Mich. 1994) ("Standard deviations are a measurement of the probability that a result is a random deviation from the predicted result."); Ottaviani v. State U. of N.Y. at New Paltz, 875 F.2d 365, 371 (2d Cir. 1989) ("A finding of two standard deviations corresponds approximately to a one in twenty, or five percent, chance that a disparity is merely a random deviation from the norm . . . ."); Smith v. City of Boston, 144 F. Supp. 3d 177, 192 (D. Mass., 2015) (stating that if the "mean test scores of minority candidates were located 1.96 standard deviations away from the overall mean, there would be only a 5% probability that such difference was due to chance"); *see also* Browne, *Beyond Damned Lies*, *supra* note 5, at 491–92 nn. 46–47 (citing additional cases). *Cf.* Tagatz v. Marquette, 861 F.2d 1040, 1044 (7th Cir. 1988) (In *Tagatz*, Judge Posner hedged about the meaning of the *p*-value: "One of Dr. Tagatz's tables comparing salary raises for Catholic and for non-Catholic faculty

levels (also called 'p-values') are simply the probability that the observed disparity is random–the result of chance fluctuation or distribution. For example, a 0.05 probability level means that . . . there is only a five percent chance that the disparity is random."[32] Similarly, the Fifth Circuit has stated that results that are statistically significant at the .05 level mean that "it could be said with a 95% certainty that the outcome was not merely a fluke."[33] Sometimes the significance level is equated to the risk of "Type I error" (false positives), a form of reasoning that also entails the transposition fallacy.[34] But to say that the *p*-value represents the likelihood that the result is a "fluke" or the likelihood of a Type I error is to say that it represents the probability that the null hypothesis is true, which it simply does not.

Some courts have gone even farther, stating not just that the *p*-value establishes the probability that chance was the cause of a disparity, but also that it strongly bears on (or even equals) the probability that the employer did not discriminate. For example, in *Adams v. Ameritech Services, Inc.*, the Seventh Circuit stated: "If the observed percentage of African-American hires is only 20%, then the statistician will compute the 'standard deviation' from the expected norm and indicate *how likely it is that race played no part in the decisionmaking*."[35] Similarly, in *Matthews v. Waukesha County*, the court stated that "[s]tatisticians typically calculate the standard deviation from the norm to determine *the likelihood that race played no role in a decision.*"[36] And in *Dobbs-Weinstein v. Vanderbilt University*, the court stated that "[t]wo standard deviations translates, approximately, into a one in twenty, or five percent (.05), chance that a particular disparity is due to chance, *and not to a particular factor such as sex.*"[37] Even expert witnesses testify in the same

---

in the school of education reveals a difference (favoring the Catholics) that is significant at the .0048 level. This means [*speaking very crudely* (*see* Kruskal, *Tests of Significance*, *in* 2 Int'l Encyclopedia of Statistics 944, 957 (1978))] that there is a probability of less than 5 in 1,000 that the difference is due to chance"). "Speaking very crudely," indeed. *Id.*

[32] Stagi v. AMTRAK 391 Fed. App'x 133, 137 (3d Cir. 2010); *see also* Tabor v. Hilti, Inc., 703 F.3d 1206, 1223 (10th Cir. 2013) ("Statistical significance measures the likelihood that the disparity between groups is random, i.e., solely the result of chance.").

[33] Rivera v. City of Wichita Falls, 665 F.2d 531, 545 n.22 (5th Cir. 1982); *see also* Anderson v. Douglas & Lomason Co., 26 F.3d 1277, 1308 n.16 (5th Cir. 1994) ("There is just a little more than one chance in a thousand that D&L's hiring during the Grizzard years happened by chance–2.4 standard deviations."). In the District Court's massive opinion in *Vuyanich v. Republic Nat'l Bank of Dallas*, 505 F. Supp. 224, 347–48 (N.D. Tex. 1980), *vacated*, 723 F.2d 1195 (5th Cir. 1984), *cert. denied*, 469 U.S. 1073 (1984), the court noted in passing that "[a] test of statistical significance does not determine the probability that any particular result in fact occurred by chance," but there is no indication that this observation played a role in the court's analysis of the statistical evidence, since the court also referred to disparities that were significant at the 5-percent level as "apparently discriminatory results," without explaining why the results would appear discriminatory if the 5-percent level is not the probability of chance occurrence.

[34] *See, e.g.*, Richard Goldstein, *Two Types of Statistical Errors in Employment Discrimination Cases*, 26 JURIMETRICS J. 32, 38 (1985) (stating that "one could set the significance level (probability of Type 1 error) at .05"); ZIMMER ET AL., 8th ed., *supra* note 15, at 135 ("Setting the level of significance at 0.05 means that a Type I error is made in only 5 percent of the cases, that is, 5 in 100 times.").

[35] 231 F.3d at 424 (emphasis added).

[36] 937 F. Supp. 2d 975, 986 (E.D. Wis. 2013) (emphasis added).

[37] 1 F. Supp. 2d 783, 803–04 (M.D. Tenn. 1998) (emphasis added); *see also* Capaci v. Katz & Besthoff, Inc., 711 F.2d 647, 652 (5th Cir. 1983) (characterizing the *p*-value as the "*probability of unbiased*

vein.[38] Some courts will describe the meaning of the *p*-value both correctly and incorrectly in the same passage, not realizing that they are saying two different things.[39] Although the focus of this article is the prevalence of the transposition fallacy in employment discrimination cases, the fallacy is a staple of other civil cases as well.[40]

---

hiring" (emphasis added)), *cert. denied*, 466 U.S. 927 (1984); Ivy v. Meridian Coca-Cola Bottling Co., 641 F. Supp. 157, 165 (S.D. Miss. 1986) (stating that "a fluctuation of two or three standard deviations indicates that the result is caused by discriminatory intent rather than chance"); Hameed v. Iron Workers Local 396, 637 F.2d 506, 513 (8th Cir. 1980) (stating that "[i]f tests of statistical significance eliminate chance as a likely explanation for the differential pass rates, courts will presume that *the disparate pass rates are attributable to racially discriminatory selection criteria*" (emphasis added)); Kassman v. KPMG LLP, 2014 U.S. Dist. LEXIS 93022, at *8 (S.D.N.Y. July 8, 2014) (stating that sex disparities "are statistically significant at more than eleven standard deviations, *meaning that the probability that KPMG's compensation could be gender neutral is less than one in one hundred million*" (emphasis added)); EEOC v. American Nat'l Bank, 652 F.2d 1176, 1191 (4th Cir. 1981) (stating that "[t]o the extent the probability of chance is shown to be quite small, *the legal inference of discrimination* based upon a rough legal assessment that disparities are manifestly 'gross' or 'substantial' *is thus 'scientifically' confirmed*") (emphasis added).

[38] *See, e.g.*, Declaration of Jie Chen, Ph.D., at *6, Lang v. Kansas City Power & Light Co., No. 99-0463-CV-W-SOW, 1999 WL 34767779 (W.D. Mo. 1999) ("Because the computed value of the chi-square is larger than the critical value, and the computed *p*-value of this test (0.000) is far less than the significance level (.05), we must conclude that the variables of race and job category are dependent. *In other words, KCPL is using race as a factor* in deciding which employee to place in a managerial position." (emphasis added)); Report or Affidavit of Chieh-Chen Bowen, Ph.D., at *2, Siegel v. Inverness Med. Innovations, Inc., No. 09CV01791, 2010 WL 2798170 (N.D. Ohio Apr. 7, 2010) ("The probability that this data pattern happened by chance (*i.e. the RIF decisions were made without considering employees' age*) was only .036 which was a very low probability event" (emphasis added)); Affidavit of Frank B. Martin, Ph.D., at *25, Thomforde v. IBM Corp., No. 02-CV-4817 JNE/FLN, 2006 WL 6578917 (D. Minn. Mar. 8, 2006) ("This is highly statistically significant with a p-value of less than .00001. What this means is *the probability that this occurred as a result of a drawing blind to age is less than one in 100,000.*" (emphasis added)); Report or Affidavit of Dr. Patricia L. Pacey, Ph.D., Camara v. Matheson Trucking, No. 12-cv-03040-CMA-CBS, 2013 WL 10236727 (D. Colo. Sept. 27, 2013) ("The statistical test result is presented as a probability value ("p-value"), i.e., the probability that the observed difference between the average black and non-black hours (or, African and non-African hours) is a random event (i.e., due to chance) *vis-à-vis disparate treatment*" (emphasis added)); Expert Report of Burt S. Barnow, Ph.D., EEOC v. AutoZone, Inc., No. 00CV02923, 2005 WL 6581866 (W.D. Tenn. Jan. 31, 2005), ("I report the probability AutoZone would have hired the number of Blacks or women actually observed *if the firm was hiring fairly* . . . ." (emphasis added)).

[39] For example, one district court described the meaning of the *p*-value as follows:

> When a statistician computes the *p*-value for any set of data, he or she is determining the probability of getting, just by chance, test data as extreme as the actual data obtained, given that the null hypothesis is true . . . . A "null hypothesis" is the hypothesis that there is no difference between two groups from which samples are drawn. For example, the null hypothesis in this case would be that there is no difference between Indiana Bell employees who are under forty and those forty or more in terms of the criteria used to select them for termination. *Thus, if the selection rates found in samples of the two age groups at Indiana Bell are not the same, then the* p*-value would give the probability that this data resulted from "the luck of the draw."* Large p-values are consistent with the null hypothesis, and small *p*-values undermine the hypothesis . . . . However, *p* does not express anything about the accuracy of the null hypothesis, or the probability that it is true. Rather, it is computed by *assuming* the hypothesis is true.

Allard v. Indiana Bell Telephone Co., 1 F. Supp. 2d 898, 907 (S.D. Ind. 1998) (first emphasis added). Except for the italicized sentence, the court correctly stated the meaning of the *p*-value, and, in fact, specifically disclaims the fallacy in the last two sentences. Right in the middle, however, is the italicized sentence clearly embodying the transposition fallacy, a sentence that is explicitly contradicted just two sentences later.

[40] See, e.g., In re Avandia Marketing, Sales Practices and Prod. Liab. Litig., 2011 WL 13576, at *12 (E.D. Pa. 2011) ("The p-value was .08, which means that there is a 92% likelihood that the difference

Judges seem to come by their misconceptions about the meaning of *p*-values honestly, because expert witnesses routinely mislead them, whether intentionally or not, probably more the latter than the former. Expert reports and testimony embodying the transposition fallacy are legion.[41] In their

---

between the two groups was not the result of mere chance."); In re Phenylpropanolamine (PPA) Prod. Liab. Litig., 289 F. Supp. 2d 1230, 1236 n.1 (W.D. Wash. 2003) ("P-values measure the probability that the reported association was due to chance . . . ."); Novo Nordisk A/S v. Caraco Pharmaceut Lab., 775 F. Supp. 2d 985, 1018 n.20 (E.D. Mich., 2011) ("The p-value is a value that statisticians use to show the uncertainty in the results of a study. Values of 0.05 or less mean that there is 5 percent or less likelihood that the outcome is the result of pure chance."); Novo Nordisk A/S v. Caraco Pharma. Lab., 719 F.3d 1346, 1350 n.3 (Fed. Cir. 2013) ("The p-value is a value that statisticians use to show the level of uncertainty in a study's results. A p-value is 'statistically significant' if it is 0.05 or less, which indicates that there is 5% or less likelihood that the outcome was the result of pure chance (citation omitted)); Acorda Therapeutics, Inc. v. Roxane Laboratories, Inc., No. CV 14-882-LPS, 2017 WL 1199767, at *37 n.10 (D. Del., Mar. 31, 2017) (citing expert witness's testimony that a p-value of 0.14 indicates "a 14% likelihood that the measured result was due to chance"); In re Testosterone Replacement Therapy Products Liability Litigation, No. 14-C-1748, MDL No. 2545, 2017 WL 1833173, at *4 (N.D. Ill., May 8, 2017) (noting that a result "would be considered statistically significant if there is a 95% probability, also expressed as a "p-value" of <0.05, that the observed association is not the product of chance"); In re PTC Therapeutics, Inc. Securities Litigation, Civ. No. 16-1124 (KM) (MAH), 2017 WL 3705801, at *16 n.4 (D. N.J., Aug. 28, 2017) (stating that "a p-value of 0.05 means that there is a 5% likelihood that an occurrence was the result of chance alone"); In re Urethane Antitrust Litigation, 166 F. Supp. 3d 501, 508 (D.N.J. 2016) (equating the *p*-value with the probability of a false positive).

[41] *See, e.g.*, Report or Affidavit of Kyle E. Brink, Ph.D., Smith v. City of Jacksonville, No. 3:11 cv 00345-TJC, 2013 WL 10154736 (M.D. Fla. Oct. 7, 2013) ("[I]f the p-value resulting from the statistical test is less than .05, then there is less than a 5% probability that the difference is due to chance. Conversely, you can conclude that there is a 95% probability that the difference is not due to chance."); Statistical Analysis and Report of Richard F. Tonowski, EEOC v. TEPRO, No. 4:12-cv-00075, 2014 WL 7778496 (E.D. Tenn. May 9, 2014) ("The t-test at the bottom of Exhibit 2 establishes that the probability of an age difference between the groups having occurred by chance alone is less than 1 in 10,000."); Report or Affidavit of Dr. Patricia L. Pacey, Ph.D., Camara v. Matheson Trucking, No. 12-cv-03040-CMA-CBS, 2013 WL 10236727 (D. Colo. Sept. 27, 2013) ("The statistical test result is presented as a probability value ("p-value"), i.e., the probability that the observed difference between the average black and non-black hours (or, African and non-African hours) is a random event (i.e., due to chance) vis-à-vis disparate treatment."); Analysis of Hiring and Economic Damages of Dr. Robert Martin LaJeunesse, M.A., Ph.D., B.S.B.A., EEOC v. FAPS, Inc., No. 310CV03095, 2013 WL 10104409 (D.N.J. Aug. 15, 2013) ("The Z-score for this period of -9.16 indicates that the aggregate shortfall is statistically significant, and that there is effectively zero probability that this outcome could have occurred by chance (p-value 0)."); Discrimination Analysis Addendum of Dr. Mark S. McNulty, B.S., Ph.D., EEOC v. JBS USA, LLC, No. 110CV02103, 2013 WL 10871078 (D. Colo. Jul. 23, 2013) ("The statistical test result is presented as a probability value, i.e., the probability that the observed differences in the rates at which disciplinary terminations are taken by Swift is due to chance."); Declaration of Dwight D. Steward, Ph.D., in Support of Plaintiff's Reply Brief re Motion for Class Certification, Ordonez v. Radio Shack, No. CV 10-07060 CAS (JCGx), 2012 WL 8964096 (C.D. Cal. Nov. 19, 2012) ("The probability that the difference was generated by random chance is referred to as a probability value or p-value for short."); Report or Affidavit of Robert M. LaJeunesse, PhD., EEOC v. Presrite, No. 11-cv-00260, 2012 WL 7075122 (N.D. Ohio June 28, 2012) ("Although the p-values for the pooled result shown in Table 14 are greater than zero, they are well below the critical significance level of 5 percent. The interpretation of the small p-values is that the probability that this hiring outcome occurred by chance is far below five, or even one, percent."); First Expert Report of Dr. Steven Wolfson, Caldwell v. University of Houston, No. 11CV02014, 2011 WL 8190124 (S.D. Tex. Nov. 29, 2011) ("At this level of statistical significance, we can be 95% sure of our result, or a one in twenty (1:20) chance of being incorrect, the so-called level of confidence."); Declaration and Expert Opinion of Dr. Venkareddy Chennareddy, Moses v. Dodaro, No. 06-1712 (EGS), 2011 WL 6442182 (D.D.C. Sept. 1, 2011) ("[T]he conclusion that those over 50 years old compared to those in 40-50 years old were discriminately impacted based on age stereotype-bias, being wrong is almost zero or has a chance of only 2.2 out of 1000."); Expert Report of Stan V. Smith, Ph.D., Bolden v. Walsh Group, No. 106CV04104, 2010 WL 8585380 (N.D. Ill. Nov. 1, 2010) ("Moreover, this difference is statistically significant at the .01% level . . . meaning that there is a less than 1 in 10,000 chance that this difference can be explained by random chance."); Report or Affidavit of Chieh-Chen Bowen, Ph.D., Siegel v. Inverness Medical Innovations, Inc., No. 09CV01791, 2010 WL 2798170 (N.D. Ohio Apr. 7, 2010) ("The probability that this data pattern happened by chance (i.e. the RIF decisions were made without considering employees' age) was only .036 which was a very low probability event."); Videotaped Deposition of Nitin Paranjpe, Ph.D., Allen v. Sears Roebuck & Co., No. 2:07-CV-11706,

opinions, many courts claim to be repeating what they have been told by experts,[42] although one never knows for sure whether a paraphrase of expert

---

2010 WL 7023485 (E.D. Mich. Apr. 1, 2010) ("And my statistical test is saying what's the p-value, what's the probability that those disparities we saw purely happened accidentally by randomness. It's very small, .022. So we rule out chance as an explanation for the disparity."); Corrections to the Reply Report of Louis R. Lanier, EEOC v. Bloomberg, No. 07 CV 8383 (LAP) (HP), 2010 WL 2103147 (S.D.N.Y. Feb. 8, 2010) ("Typically, in the economics professional literature, standard deviations of approximately 2 [sic] or greater in absolute value (1.96, to be more precise) are considered statistically significant, representing a five percent level of probability (1 in 20) that the tested result occurred by chance."); Declaration of Kyle Brink, Ph.D., Howe v. City of Akron, No. 5:06 CV 2779, 2008 WL 8568234 (N.D. Ohio Sept. 6, 2008) (If the p-value resulting from the statistical test is less than .05, then there is less than a 5% probability that the difference is due to chance. Conversely, you can conclude that there is a 95% probability that the difference is not due to chance.); Report or Affidavit of James R. Lackritz, Ph.D., Terry v. City of San Diego, No. 06CV01459, 2008 WL 8698107 (S.D. Cal. Sept. 22, 2008) ("The p-value of .000 means the probability of seeing this female/male distribution between the LGI and higher classifications is .000 when rounded to the third decimal place, or virtually impossible. Any probability lower than .05 is considered to be too low to believe that the difference in these distributions are merely due to random chance."); Report or Affidavit of Stan V. Smith, Ph.D., Derrico v. MGM Grand Detroit, LLC, No. 03CV73133, 2007 WL 6519719 (E.D. Mich. Oct. 15, 2007) ("[T]he chance that so many Black employees were demoted based solely on a random selection is approximately three in one billion."); Affidavit of Kurt V. Krueger, Ph.D., Johnson v. United States Beef Corp., No. 04:04-CV-00963-FJG, 2006 WL 3861830 (W.D. Mo. Nov. 7, 2006) ("The Fisher's Exact p-value is p= . . . .05 which indicates a probability value of 5% that the differences in racial composition between the Arby's store and that of the comparable labor force in Lee's Summit occurred by chance."); Affidavit of Frank B. Martin, Ph.D., Thornforde v. IBM Corp., No. 02-CV-4817 JNE/FLN, 2006 WL 6578917 (D. Minn. Mar. 8, 2006) ("This is highly statistically significant with a p-value of less than .00001. What this means is the probability that this occurred as a result of a drawing blind to age is less than one in 100,000."); Report or Affidavit of J. Michael Hardin, Ph.D., U.S. v. Jefferson County, No. CV-75-S-0666-S, 2005 WL 6000850 (N.D. Ala. Jan. 24, 2005) ("Using this convention of rejecting the null hypothesis when it is less than a pre-set value such as .05 leads to a decision rule that assures the analyst that the mistake of rejecting the null hypothesis when it is actually true (called a Type I error) will only occur with probability of .05."); Expert Report of Dr. Kathleen K. Lundquist, U.S. v. Jefferson County, No. CV-75-S-0666-S, 2005 WL 6000852 (N.D. Ala. Feb. 7, 2005) ("Using the Fisher's Exact Probability Test results in a probability value (*i.e.*, p-value) which indicates the likelihood that the observed differences in selection rates occurred by chance alone."); Expert Report of Ali Saad, Ph.D. (defense expert), Sepulveda v. Wal-Mart Stores, Inc., No. 204CV1003, 2004 WL 5389362 (C.D. Cal. 2004) ("[A .05] level of statistical significance is equivalent to saying that there is less than a one in twenty chance that the observed relationship is due to chance."); Expert Report of James W. Meeker, J.D., Ph.D., Deja Vu-Toledo, Inc. v. City of Toledo, No. 3:03CV7245, 2004 WL 5740201 (N.D. Ohio July 20, 2004) ("This error rate states we are 95 per cent confident the effect detected is a real effect, not an artifact of the sample."); Expert Trial Transcript of Mark Berman, NAACP v. Florida Department of Corrections, No. 5:00-cv-100-OC-10GRJ, 2003 WL 24296516 (M.D. Fla. Nov. 12, 2003) ("[W]ith the 'P' value of.02, I can say that I'm 98 percent sure that the result I came up with did not happen as a consequence of chance."); Brief of the Equal Employment Opportunity Commission as Amicus Curiae in Support of Plaintiff's Appellants and Reversal, Cooper v. Southern Co., 390 F.3d 695 (11th Cir. 2004) (quoting expert as testifying with respect to a racial difference in promotions, "The probability that this racial difference could have occurred by chance is less than 26 in ten trillion"); Declaration of Jie Chen, Ph.D., Lang v. Kansas City Power & Light Co., No. 99-0463-CV-W-SOW, 1999 WL 34767779 (W.D. Mo. 1999) ("My standard for choosing the significance level for hypothesis testing of the data is .05, This means that the results of this testing have a 95% probability of being accurate, and only a 5% chance of error"); Opinions and Report of Bill Luker, Jr., Ph.D., Bolton v. Lear Seating Corp., No. 97-C-392-C, 1998 WL 35074551 (W.D. Wis. June 1, 1998) ("'[*P*] = .0026' indicates that the probability that this difference arose merely by chance is less than 1 percent, and so on through the rest of the tests."); Declaration of Jie Chen, Ph.D., Turner v. Torotel, Inc., No. 96-0646-CV-W-5, 1996 WL 34388584 (W.D. Mo., Oct. 29, 1996) ("My standard for choosing the significance level for hypothesis testing of the Torotel data is .05. This means that the results of this testing have a 95% probability of being accurate, and only a 5% chance of error."); *see also* Report of Richard Drogin, Ph.D., McClain v. Lufkin Industries, No. 97CV00063, 2003 WL 25859212 (E.D. Tex. June 16, 2003) ("If the race coefficient is statistically significant, the disparity cannot be explained by chance variation and the coefficient indicates a real difference between black and non-black pay rates.").

[42] *See* Blum v. Witco Chem. Corp., 829 F.2d 367, 371 (3d Cir. 1987) ("Additionally, plaintiffs produced a statistical expert who testified that the probability that the disparate retention rate was due to some random factor unrelated to age was .0084."); Moore v. McGraw Edison Co., 804 F.2d 1026, 1031 (8th Cir. 1986) ("[Plaintiffs' expert] testified that the chances were two in one thousand that age was not a factor in the terminations."); EEOC v. Olson's Dairy Queens, Inc., 989 F.2d 165, 167 (5th Cir. 1993) ("Dr. Straszheim concluded that the likelihood that [the] observed hiring patterns resulted from truly race-neutral

testimony is faithful to the original, and it is easy to inadvertently paraphrase a statement that correctly states the meaning of the *p*-value into a statement that embodies the transposition fallacy. The ambiguity about the origin of the misunderstanding is demonstrated in *Brackett v. Civil Service Commission*,[43] in which the court quoted the plaintiff's statistician as follows:

> Statisticians measure how far away an actual outcome is from what's expected in statistical units called standard deviations. A standard deviation of two approximately corresponds to a probability of occurrence of five percent.

Putting aside the nonsense about "a standard deviation of two"—when (one hopes) the expert meant "two standard deviations," a very different animal—it is not clear whether the testimony embodies the fallacy because the statement refers vaguely to "a probability of occurrence" but doesn't say the probability of *what* occurrence. The statement could mean, correctly, that the probability that a nondiscriminating employer chosen at random from a pool of nondiscriminating employers would have a disparity that great or greater is five percent, or it could mean, incorrectly, that the probability that an employer with this great a disparity was selecting randomly is five percent. In any event, the court immediately followed the expert's quotation with a quotation from another case, which clearly rests on the fallacy: "Two standard deviations is normally enough to show that it is extremely unlikely (that is, there is less than a 5% probability) that the disparity is due to chance, giving rise to a reasonable inference that the hiring was not race-neutral."[44] The court later paraphrased additional testimony by the same expert as asserting "that a shortfall of 16.2 women was significant and corresponded to a statistical disparity of 3.14 standard deviations, meaning that there was approximately a one in 600 probability that such occurrence had happened by chance,"[45] another statement that embodies the fallacy. Perhaps the expert's testimony itself did not explicitly embody the transposition fallacy, but it does not seem that the expert worked very hard to prevent the court from lapsing into it.

Another example of the difficulty that courts have keeping the matter straight can be found in the District Court's opinion in *Bazile v. City of Houston*.[46] The court quoted the plaintiff's expert's report as follows:

> Statistical tests produce a probability value . . . that determines or estimates the probability of obtaining the sample result

---

hiring practices was less than one chance in ten thousand."); People v. Washington, 179 P.3d 153, 163 (Colo. App. 2007) (discussing the defendant's argument that there was discrimination in jury selection, stating that defendant's expert "opined that the probability of the .3% disparity being attributable to chance was .008% (or eight chances in 100,000")).

[43] 850 N.E.2d 533, 543 n.14 (Mass. 2006).

[44] *Id.* (*quoting* Adams v. Ameritech Servs., Inc., 231 F.3d 414, 424 (7th Cir. 2000)).

[45] *Id.* at 549.

[46] 858 F. Supp. 2d 718 (S.D. Tex., 2012).

> assuming there were no differences in the population. . . . For example, if an alpha level of .05 is chosen and the [probability value] resulting from the statistical test is less than .05, then there is less than a 5% probability that the difference is due to chance (i.e., there is less than a 5% probability of making a Type I error) and we say the result is statistically significant. Conversely, you can conclude that there is a 95% probability that the difference is not due to chance.[47]

The first sentence is a correct statement of the meaning of the *p*-value: the likelihood of the result *if* the null hypothesis is true. The remainder of the quotation reflects the transposition fallacy: the *p*-value is the probability that the result is due to chance (and therefore the probability *that* the null hypothesis is true).

The United States Supreme Court has not been spared such arguments, although it has not engaged in the fallacy itself. In *Matrixx Initiatives, Inc., v. Siracusano*,[48] a brief filed on behalf of a pair of self-described "statistics experts" asserted the following:

> The 5 percent significance rule insists on 19 to 1 odds that the measured effect is real. There is, however, a practical need to keep wide latitude in the odds of uncovering a real effect, which would therefore eschew any bright-line standard of significance. Suppose that a p-value for a particular test comes in at 9 percent. Should this p-value be considered "insignificant" in practical, human, or economic terms? We respectfully answer, "No." For a p-value of .09, the odds of observing the AER [adverse event report] is 91 percent divided by 9 percent. Put differently, there are 10-to-1 odds that the adverse effect is "real" (or about a 1 in 10 chance that it is not). Odds of 10-to-1 certainly deserve the attention of responsible parties if the effect in question is a terrible event.[49]

Fortunately, the Supreme Court did not specifically address that point in its opinion, although it did make a couple of unfortunate observations about

---

[47] *Id.* at 738 (alteration in original).

[48] 563 U.S. 27 (2011). *Matrixx* presented the question whether plaintiffs can state a claim for securities fraud based upon a pharmaceutical company's failure to disclose adverse event reports even "if the reports do not disclose a statistically significant number of adverse events." *Id.* at 29 (holding that plaintiffs could state such a claim. In fact, however, "[b]ecause case reports are just a series of anecdotes, it is not immediately obvious how they could be statistically significant." David H. Kaye, *The Transposition Fallacy in* Matrixx Initiatives, Inc. v. Siracusano*: Part I*, FORENSIC SCI., STATS. & LAW (2011), http://for-sci-law.blogspot.com/2011/08/one-might-expect-to-hear-phrases-like.html (last visited May 1, 2017).

[49] Brief of Amici Curiae Statistics Experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in Support of Respondents, at *18, Matrixx Initiatives, Inc. v. Siracusano, 2010 WL 4657930.

statistics and causation.[50]

Plaintiffs' lawyers have an obvious incentive to perpetuate the transposition fallacy.[51] One might suppose that defense lawyers and their experts would be standing ready to disabuse courts of their embrace of the fallacy, but such does not seem to be the case.[52] I have been unable to find a single reported case in which such an objection was mentioned in an opinion.[53] Indeed, it is not clear that defense experts are any less likely to lapse into the fallacy than anyone else.[54]

Surely, one might think, at least the fallacy must be well-understood by all scholars of law and statistics. There again, the answer is no. One finds

---

[50] *See* Nathan A. Schachtman, *Statistical Evidence in Products Liability Litigation*, *in* PRODUCT LIABILITY LITIGATION: CURRENT LAW, STRATEGIES AND BEST PRACTICES (Stephanie A. Scharf, Lise T. Spacapan, Traci M. Braun, & Sarah R. Marmor, eds.) 30A-9 - A-12 (2014) (noting that the Court in dictum made comments about the lack of need for statistically significant evidence of causation in pharmaceutical cases that are inconsistent with current law.

[51] *See* Steven Rotman, *Don't Know Much About Epidemiology? Gain a Strategic Advantage in Pharmaceutical Litigation by Boning up on Epidemiology*, 43 AM. ASS'N FOR JUSTICE, at 35 (2007) (stating that *p*-values "measure the probability that a reported association between a drug and condition was due to chance. A P value of .05, which is generally considered the standard for statistical significance, means there is a 5 percent probability the association was due to chance"); Christine E. Webber, *A Plaintiff's Perspective on Some Evidentiary Issues and Jury Instructions in Employment Discrimination Litigation*, ABA BUS. L. COURSE MATERIALS J. (2008) (providing a "sample jury instruction" stating, "[a]t 1.96 standard deviation [sic], there is no more than a 5% chance that the pay differences found would arise solely by chance. In other words, one can be 95% certain that there really is a difference in how the groups being studied are compensated"[;] the instruction also tells the jury, "No minimum degree of disparity or statistical significance must be met to establish a claim of discrimination.").

[52] *See* Bruce R. Parker & Anthony F. Vittoria, *Debunking Junk Science: Techniques for Effective Use of Biostatistics*, 65 DEF. COUNS. J. 33, 44 (2002) ("[A] P value of .01 means the researcher can be 99 percent sure that the result was not due to chance.").

[53] There are a few unreported cases in which the expert mentioned the transposition-fallacy argument. *See* Report of James T. McClave, Ph.D., U.S. v. City of Jacksonville, No. 3:12-cv-451-J-32 MCR, 2014 WL 7778588 (M.D. Fla. Jul. 14, 2014) ("Drs. Brink's and Siskin's conclusory statements about the probability the disparity is due to chance is known as the 'fallacy of the transposed conditional' . . . . Instead of correctly concluding that a statistically significant result means that the disparity would be unlikely to have occurred . . . assuming that chance is the cause, they concluded that they know the probability of cause being the chance having observed a particular disparity. In fact, the test result does [not] provide that probability."); Amended Expert Report of Richard McCleary, Ph.D., McGuire v. City of Montgomery, No. 2:11-CV-1027-WKW, 2014 WL 8096111 (M.D. Ala. Jan. 7, 2014) ("Finally, in §5.3, I discuss Bayesian inverse probabilities. Misinterpretations of statistical significance usually confuse the significance level (or p-value) of a hypothesis test with the Bayesian inverse probability of the hypothesis. Since the p-value and corresponding Bayesian inverse probability are not generally equal, this is a serious error."); Videotaped Deposition of Franklin M. Fisher, U.S. ex rel. Tyson v. Amerigroup Ill., Inc., No. 04 C 6074, 2006 WL 3039354 (N.D. Ill. Sept. 12, 2006) ("But the P value is not the probability that the null hypothesis is false, nor is it the probability that the null hypothesis is true. It's the probability of obtaining the results you've obtained if the null hypothesis is true.") (False Claims Act case); Affidavit of Joseph B. Kadane, Ph.D., Hall v. Best Buy Co., No. 04-4812 (MJD/JGL), 2006 WL 6610712 (D. Minn. June 29, 2006) ("A well-recognized danger of the use of tests of significance is that they invite the unwary reader, juror or judge to confuse the probability of the data as or more extreme than that observed if the null hypothesis were true (which is a legitimate conclusion from a test of significance) with the probability that the null hypothesis is true, given the data (which is not a legitimate conclusion from this method).")).

[54] *See, e.g.*, Clark v. Commonwealth of Pennsylvania, 885 F. Supp. 694, 707 (E.D. Pa. 1995) (stating that "[t]he parties further agree that under the Fisher's Exact Test, there was a .0194 probability that the disparity in selection rates between black and white employees who had passed the exam was due to random chance . . . "); Delgado-O'Neil v. City of Minneapolis, 745 F. Supp. 2d 894, 911 (D. Minn. 2010) (stating that "both experts agreed that the standard for finding statistical significance is a p-value (probability value) of .05 or less (a 1 in 20 chance that the event occurred by chance)").

ample occurrences of the fallacy in both books[55] and journal articles[56] dealing with employment discrimination and statistics.

Perhaps one should not condemn too strongly those in the legal field who fail to understand techniques that are not, for most of them, central to their specialty. Studies of academic psychologists, for many of whom null-hypothesis significance testing is their bread and butter, also find widespread misunderstanding of the meaning of *p*-values, such that large percentages of students and researchers (even those who teach statistics) cannot correctly answer basic questions on the subject.[57] As psychologist Geoff Cumming asks, "If a technique is not even understood correctly by its teachers, what

---

[55] JOEL FRIEDMAN, THE LAW OF EMPLOYMENT DISCRIMINATION: CASES AND MATERIALS 342 (9th Ed. 2013) ("[T]he standard deviation is a way to calculate the likelihood that chance is responsible for the difference between a predicted result and an actual result."); SAMUEL ESTREICHER & MICHAEL C. HARPER, CASES AND MATERIALS ON EMPLOYMENT DISCRIMINATION LAW 67, 67 (4th Ed. 2012) ("[T]he binomial model provides a statistical means for determining whether the observed outcomes are likely to be the product of chance (the "null hypothesis)."); MACK A. PLAYER, EMPLOYMENT DISCRIMINATION LAW 347, 344–49 (1987) (stating that if the *p*-value is .04, "we are 96% sure that the result did not occur by chance . . . ."); *cf.* DIANNE KRUMM, PSYCHOLOGY AT WORK: AN INTRODUCTION TO INDUSTRIAL/ORGANIZATIONAL PSYCHOLOGY 546, 546 (2001) (stating that "[t]he .05 significance level means that the probability is 95 percent that the results of an experiment are caused by the experimental manipulations and are not due to chance"); DAVID BALDUS & JAMES COLE, STATISTICAL PROOF OF DISCRIMINATION § 9.02, at 290 (1980) ("A finding that a disparity is statistically significant at the 0.05 or 0.01 level means that there is a 5 per cent. or 1 per cent. probability, respectively, that the disparity is due to chance"); *see also* Browne, *Beyond Damned Lies*, *supra* note 5, at 493–94 n. 49 (collecting sources).

[56] Tristin K. Green, *Discrimination in Workplace Dynamics: Toward a Structural Account of Disparate Treatment Theory*, 38 HARV. CIV. RTS.-CIV. LIB. L. REV. 91, 121 n.130 (2003) ("Statistical significance in employment discrimination suits is often determined using a binomial distribution analysis, which measures the probability that a certain outcome is due to chance."); Jennifer L. Peresie, *Toward a Coherent Test for Disparate Impact Discrimination*, 84 IND. L.J. 773, 774 (2009) ("Under statistical significance tests, a disparity is actionable when we can be confident at a specified level–generally ninety-five percent–that the observed disparity is not due to random chance."); Cheryl I. Harris & Kimberly West-Faulcon, *Reading* Ricci*: Whitening Discrimination, Racing Test Fairness*, 58 U.C.L.A. L. REV. 73, 137 n.237 (2010) ("The *p*-value is the chance or likelihood–'significance probability'–that the observed racial disparity happened by chance (without regard to race)."); Michael I. Meyerson & William Meyerson, *Significant Statistics: The Unwitting Policy Making of Mathematically Ignorant Judges*, 37 PEPP. L. REV. 771, 824 (2010) (stating that under a .05 significance level, one will incorrectly condemn the innocent only five percent of the time); *id.* at 827 (stating that "[i]nnocent employers will lose only one time out of twenty . . ."); David W. Barnes, *The Significance of Quantitative Evidence in Federal Trade Commission Deceptive Advertising Cases*, 46 LAW & CONTEMP. PROBS. 25, 39 (1983) ("It is commonly accepted that a significance level of 0.05, meaning that there is at most one chance in twenty that the difference is not real but is just due to chance, is a small enough probability of error that scientists can conclude that the differences are real."); Lucinda M. Finley, *Guarding the Gate to the Courthouse: How Trial Judges Are Using Their Evidentiary Screening Role to Remake Tort Causation Rules*, 336 DEPAUL L. REV. 335, 348 n.49 (1999) ("Courts also require that the risk ratio in a study be 'statistically significant,' which is a statistical measurement of the likelihood that any detected association has occurred by chance, or is due to the exposure."). *See also* David L. Schwartz & Christopher B. Seaman, *Standards of Proof in Civil Litigation: An Experiment from Patent Law*, 26 HARV. J. LAW & TECH. 458, 460 n.187 (2013) ("Statistical significance is the probability that an observed relationship is not due to chance."); Margaret G. Farrell, Daubert v. Merrell Dow Pharmaceuticals, Inc.: *Epistemiology and Legal Process*, 15 CARDOZO L. REV. 2183, 2208–09 (1994) ("Statistical significance means that the likelihood that the test results . . . occurred by chance are less than some amount, established by convention at 5%."); Richard A. Posner, *An Economic Approach to the Law of Evidence*, 51 STAN. L. REV. 1477, 1510–11 (1999) (stating that statistical significance at the five percent level "mean[s] that the probability that the investigation would have yielded this result even if the hypothesis that it was trying to test was false is no greater than five percent").

[57] CUMMING, *supra* note 8, at 25–26.

hope is there for students and teachers who wish to use it?"[58] One might further ask, "If a technique is not even understood correctly by those for whom it is a primary tool of their discipline, what hope is there for judges and juries for whom statistical techniques are, by and large, a foreign language?"

The ubiquity of the fallacy in the cases and its prevalence in the literature should not mislead the reader into thinking that adequate cautions are not widely available. There is ample commentary in the literature on this specific problem. No one has been clearer in explaining the fallacy than David Kaye:

> The court's assumption, however, that when the "probability of statistical error is less than 5%," the "scientific fact is at least 95% certain" exemplifies a common misunderstanding of the role of statistical tests in scientific inference. . . . The difficulty is that this interpretation of the result of the hypothesis test is wrong. The test was structured so as to retain the null hypothesis unless the chance of getting the evidence under this hypothesis fell below 5%. The test focused exclusively on the probability of the evidence given the null hypothesis. Nothing was said about the probability of the hypothesis in the light of the experimental evidence. It may be tempting to call the probability of 0.055 the chance of a coincidence, and to say that the probability of something other than a coincidence–of foul play– must be what is left over, namely 0.945. But this only shows that one can "prove" anything with words.[59]

Similarly, *The Reference Manual on Scientific Evidence*, provided free by the Federal Judicial Center to all members of the federal judiciary (and available without charge to all online), is also clear:

> Because *p* is calculated by assuming that the null hypothesis is correct, *p* does not give the chance that the null is true. The *p*-value merely gives the chance of getting evidence against the null hypothesis as strong as or stronger than the evidence at hand. Chance affects the data, not the hypothesis. According to the frequency theory of statistics, there is no meaningful way to assign a numerical probability to the null hypothesis. The correct interpretation of the *p*-value can

---

[58] *Id.* at 26.

[59] David H. Kaye, *Statistical Significance and the Burden of Persuasion*, 46 LAW & CONTEMP. PROBS. 13, 21–2 (Autumn 1983); *see also* Robert Follett & Finis Welch, *Testing for Discrimination in Employment Practices*, 46 LAW & CONTEMP. PROBS. 171, 174 (Autumn 1983) (stating that "the 0.05 rule does not say that when a difference as large as two standard deviations occurs the probability that the two groups are treated equally is 5% or less, nor does it say that the probability of unequal treatment is 95% or more.").

therefore be summarized in two lines:

> *p* is the probability of extreme data given the null hypothesis.
>
> *p* is not the probability of the null hypothesis given extreme data.[60]

These clear explanations have not, unfortunately, been widely internalized. One of the reasons for the persistence of the fallacy is that most "statistical experts" who testify are not actually statisticians. Instead, they come from fields such as psychology, sociology, or economics, and they use statistics as tools of their discipline without necessarily understanding the underlying principles.

D. Why it Matters

Why does this all matter? Isn't this just a debate about arcana of interest only to statistics nerds? The answer is that it matters a lot, because adoption of the fallacy causes courts to misunderstand what plaintiffs have proven, and their misunderstanding is reinforced by the patina of precision and objectivity seemingly conferred by mathematics.[61]

1. The Transposition Fallacy Leads to an Unwarranted Sense of Certainty

When courts reason from a *p*-value of .05 that there is only a five-percent chance that a disparity was caused by chance, they have effectively ruled out chance as the cause and expect to see evidence from the employer about what the systematic cause of the disparity was, often presuming discrimination in the absence of such evidence. As Rex Kline has observed with some understatement, "[a] researcher who mistakenly believes that low *p* values make the null hypothesis unlikely may become overly confident in the results."[62] The same goes for judges. For example, the Fourth Circuit has stated that disparities of more than two or three standard deviations

---

[60] KAYE & FREEDMAN, *supra* note 17, at 250; *see* Bent, *supra* note 28, at 820–23.

[61] A response to the above concern is that criticism of courts for committing the transposition fallacy does not take into account that the defendant may still rebut the showing that chance has been ruled out. MICHAEL J. ZIMMER, CHARLES A. SULLIVAN, & REBECCA HANNER WHITE, CASES AND MATERIALS ON EMPLOYMENT DISCRIMINATION 134, 144 (7th ed., 2008) (arguing that the defendant may rebut the showing with "[s]ufficiently strong testimony . . . that it was chance that explained the disparity") [hereinafter ZIMMER ET AL., 7th ed.]. There is much that is wrong about that argument. First, it does not take into account what it is demanding of the employer: to demonstrate that chance was responsible in the face of evidence that has already been erroneously interpreted to mean that there is a 95% probability that chance was *not* responsible, a probability level that approaches the "beyond a reasonable doubt" standard. Second, this faulty evidence has the effect of shifting the burden of proof to the employer to prove nondiscrimination, which is inconsistent with the background rule that the plaintiff bears the burden of proving discrimination. *See* Texas Dep't. of Cmty. Affairs v. Burdine, 450 U.S. 248, 254–55 (1981). Third, it is strange to argue that parties against whom faulty and misleading evidence is introduced should not complain about its admission because they can always introduce evidence in rebuttal. Taken seriously, that is an argument for repeal of all rules of evidence aimed at ensuring reliability.

[62] KLINE, *supra* note 14, at 19.

"conclusively ruled out chance as the cause of the disparity in the termination rates."[63] Many commentators also explicitly link the *p*-value with the likelihood of making a mistake, as in the statement, "[w]hen 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."[64] Instead, courts should reason that the statistics relied upon by the plaintiff would be true of thousands of non-discriminating employers who were selecting employees entirely at random.

It should be emphasized that not just five percent of employers are at risk of being unfairly tarred through operation of the transposition fallacy. Virtually *all* employers of any size, who have multiple departments, locations, job categories, etc., are likely to have some—and perhaps many—statistically significant disparities in their workforces due wholly to chance even if they do not discriminate (and many others that are due to neither chance nor discrimination).

Attempting to compute probabilities of random selection from p-values is even more inappropriate in employment cases because, in many cases, employers are subjected to litigation *precisely because they have statistical imbalances*. In research, it is generally considered unethical to simply look through a data set, find statistical relationships, and then report them as if they support some particular hypothesis. That practice, "data dredging" or "data mining," is acceptable as exploratory research that may lead to further testing *on a different population*.[65] Gerd Gigerenzer refers to this phenomenon as Feynman's Conjecture, based upon the famous physicist's reaction to a psychology researcher who, after running an experiment in which rats ran in a T-maze and finding that they did not behave as predicted, noted that the rats seemed to alternate right turns with left turns.[66] The researcher wanted Feynman to calculate the probability of getting that pattern (specifically to see if the probability was less than 5% and therefore "statistically significant"). Feynman told him that he could not select the case on the basis of the pattern revealed and then test it statistically. Instead, to test the hypothesis that rats have a tendency to alternate in the maze, it would be necessary to run the experiment on a different group of rats, not the group that generated the hypothesis. The researcher did this and came up empty.

To understand how the selection of cases would tend to increase the

---

[63] Lilly v. Harris-Teeter Supermarket, 720 F.2d 326, 336 (4th Cir. 1983), *cert. denied*, 466 U.S. 951 (1984).

[64] Louis J. Braun, *Statistics and the Law: Hypothesis Testing and its Application to Title VII Cases*, 32 HASTINGS L.J. 59, 87 (1980).

[65] To claim an effect based upon known statistical disparities is sometimes referred to as HARKing (hypothesizing after the results are known). Norbert L. Kerr, *HARKing: Hypothesizing After the Results Are Known*, 2 PERS. & SOC. PSYCHOL. REV. 196 (1998); *see* KLINE, *supra* note 14, at 73 (noting that conducting many significance tests increases the likelihood of finding associations, even implausible ones).

[66] GIGERENZER, *supra* note 14, at 602–03.

number of false positives, imagine that 100 individuals flip a coin 100 times each. Imagine also that the Department of Fair Coins is charged with detection of unfair coins (coins having something other than a 50/50 chance of flipping heads) and unfair coin flippers. Which flippers is the Department going to scrutinize? It's not going to bother with the ones that got splits of 50/50, 49/51, 48/52, 47/53 etc. Rather, it will focus on the subset with the more extreme numbers. Under the null hypothesis (that the coins were fair and the flippers were honest), one would expect that about 5% of the 100 flippers would obtain a result as extreme as, or more extreme than, a 60/40 or 40/60 split. Yet, when the Department chooses to focus on the group with imbalanced results and analyze their results statistically, a very high percentage of the chosen cases would have statistically significant results even if all of the coins were fair and all the flippers were honest. Similarly, in the employment context, one would expect the EEOC or private plaintiffs to concentrate their efforts against employers already known to have relatively large disparities and then test them for statistical significance. A large number of those *selected* employers would be expected to have statistically significant disparities even if they hire randomly. Thus, just "snooping" through data looking for significant differences—or selecting a data set that has already been observed to have statistical associations—ensures that the likelihood of Type I error is increased.[67] This practice—using the same data for selection and analysis—has been labeled "double dipping," and can "result in distorted descriptive statistics and invalid statistical inference."[68] As Ronald Coase is reported to have said, "If you torture the data long enough it will confess."[69]

2. The Transposition Fallacy Leads to Conflation of Significance Levels and Standards of Proof: Scientific versus Legal Proof

The fact that the end product of hypothesis testing is a probability and that the standard of proof is often also viewed as a probability[70] has led some courts and commentators to equate the two. After all, if social science imposes a 95%-certainty requirement, significant results should engender a

---

[67] Jonathan Taylor & Robert J. Tibshirani, *Statistical Learning and Selective Inference*, 112 PNAS 7629, 7629–30 (2015) (noting that when data have been "cherry picked"—that is, selected because of their associations—then there must be a higher bar to declare the associations significant).

[68] Nikolaus Kriegeskorte et. al., *Circular Analysis in Systems Neuroscience: The Dangers of Double Dipping*, 12 NATURE NEUROSCI. 535, 535 (2009).

[69] Gordon Tullock, *A Comment on Daniel Klein's "A Plea to Economists Who Favor Liberty,"* 27 E. ECON. J. 203, 205 (2001); *see* Harvey J. Motulsky, *Common Misconceptions about Data Analysis and Statistics*, 35 J. PHARMACOL. & EXP. THER. 200, 200 (2014) ("Keep trying until you obtain a statistically significant result or until you run out of money, time, or curiosity. The results from data collected this way cannot be interpreted at face value.").

[70] *See* Rita J. Simon & Linda Mahan, *Quantifying Burdens of Proof: A View from the Bench, the Jury, and the Classroom*, 5 LAW & SOC'Y REV. 319, 325 (1971); C.M.A. McCauliff, *Burdens of Proof: Degrees of Belief, Quanta of Evidence, or Constitutional Guarantees?*, 35 VAND. L. REV. 1293, 1293 (1982); Bradley Saxton, *How Well Do Jurors Understand Jury Instructions? A Field Test Using Real Juries and Real Trials in Wyoming*, 33 LAND & WATER L. REV. 59 (1998).

high degree of confidence in one's conclusions, far higher than the law requires (at least outside of the criminal context).[71] In an early case involving statistics, *Ethyl Corp. v. EPA*, the D.C. Circuit made explicit that confused view:

> Petitioners demand sole reliance on *scientific* facts, on evidence that reputable scientific techniques certify as certain. Typically, a scientist will not so certify evidence unless the probability of error, by standard statistical measurement, is less than 5%. That is, scientific fact is at least 95% certain.
>
> Such certainty has never characterized the judicial or the administrative process. It may be that the "beyond a reasonable doubt" standard of criminal law demands 95% certainty. But the standard of ordinary civil litigation, a preponderance of the evidence, demands only 51% certainty. A jury may weigh conflicting evidence and certify as adjudicative (although not scientific) fact that which it believes is more likely than not.[72]

A number of other courts,[73] as well as commentators,[74] have also made the

---

[71] *See* Farrell, *supra* note 56, at 2208–09 (questioning whether "the 95 percent certainty test" of science should apply in legal proceedings).

[72] Ethyl Corp. v. EPA, 541 F.2d 1, 28 n.58 (D.C. Cir.), *cert. denied*, 426 U.S. 941 (1976) (citation omitted).

[73] *See* In re Ephedra Products Liability Litigation, 393 F. Supp. 2d 181, 193 (S.D.N.Y. 2005) ("Scientific convention defines statistical significance as '$p \leq .05$,' i.e., no more than one chance in twenty of finding a false association due to sampling error. Plaintiffs, however, need only prove that causation is more-probable-than-not."); *see also id.* at 193 n.9 ("More-probable-than-not might be likened to $p<.5$, so that preponderance of the evidence is nearly ten times less significant (whatever that might mean) than the scientific standard."); Hodges v. Sec'y Dep't Health & Human Serv., 9 F.3d 958, 967 (Fed. Cir. 1993) (Newman, J., dissenting) ("Scientists as well as judges must understand 'the reality that the law requires a burden of proof, or confidence level, other than the 95 percent confidence level that is often used by scientists to reject the possibility that chance alone accounted for observed differences.'"); *id.* at 965, 965 n.4 (quoting the Report of the Carnegie Commission on Science, Technology, and Government, *Science and Technology in Judicial Decision Making* 28 (1993) (stating that "although the data may not establish a causal relationship to a medical certainty, they may nonetheless meet the more-likely-than-not standard of the law," and noting that "reasonable medical certainty" for statistical data "means that the results are statistically significant at the confidence level normally required for scientific acceptance and publication in a scientific journal–usually about 0.95."); Longmore v. Merrell Dow Pharma., Inc., 737 F. Supp. 1117, 1120 (D. Idaho 1990) ("[T]he plaintiff need only prove that it is more probably true than not that the mother's ingestion of Bendectin caused David's Poland's Syndrome. This certainly does not require a confidence level of 95%, 90% or even 80%. A cause-and-effect relationship may be deemed insignificant under stringent scientific standards, but nevertheless establish causation under legal standards.").

[74] Steven R. Weller, *Book Review: Regulating Toxic Substances: A Philosophy of Science and Law*, 6 HARV. J. L. & TECH. 435, 436, 437–38 (1993) ("[O]nly when the statistical evidence gathered from studies shows that it is more than ninety-five percent likely that a test substance causes cancer will the substance be characterized scientifically as carcinogenic. . . . [T]o determine legal causality, the plaintiff need only establish that the probability with which it is true that the substance in question causes cancer is at least [sic] fifty percent, rather than the ninety-five percent to prove scientific causality . . . ."); Braun, *supra* note 64, at 70 (stating that "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").

fundamental error of equating the *p*-value with the probability of false inculpation.

Once it is recognized that the *p*-value does not reflect the probability of making a mistake, there is no basis for viewing the preponderance of evidence and a *p*-value of .05 as two points on the same scale with the latter indicating a higher degree of certainty. Moreover, as discussed below, the "certainty" purportedly required by social science standards is anything but.

The contrast between "scientific standards" (which are thought to require near certainty) and "legal standards" (which are viewed as much more forgiving) is based upon an incorrect view of what a scientist has proved with a finding of statistical significance. A finding that results are statistically significant at the .05 level does not establish that a phenomenon exists or even that there is a 95-percent chance that it does.[75] As David Colquhon has noted, "Observation of a *P* value close to 0.05 means nothing more than 'worth another look.'"[76] He notes that "[i]f you want to avoid making a fool of yourself too often, don't regard anything bigger than $P < 0.001$ as a demonstration that you've discovered something." Ronald A. Fisher, who developed the concept of *p*-values, cautioned regarding the .05 significance level that "[a] scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance."[77] Only when an experiment is repeatedly replicated can it be viewed as an established phenomenon (still subject to later revision or even rejection).

It is thus a mistake to draw strong conclusions from a single hypothesis test. Perhaps ironically in light of the present discussion, Steven Goodman criticizes the use of hypothesis testing in medical research precisely because it fails to operate as a proper justice system would:

> Hypothesis tests are equivalent to a system of justice that is not concerned with which individual defendant is found guilty or innocent (that is, "whether each separate hypothesis is true or false") but tries instead to control the overall number of incorrect verdicts (that is, "in the long run of experience, we shall not often be wrong"). Controlling mistakes in the long run is a laudable goal, but just as our sense of justice demands that individual persons be correctly judged, scientific intuition says that we should try to draw the

---

[75] *See* Donald Berry, *Multiplicities in Cancer Research: Ubiquitous and Necessary Evils*, 104 J. NATL. CANCER INST. 1124, 1125 (stating that "[m]uch of the world acts as though statistical significance implies truth, which is not even approximately correct").

[76] David Colquhoun, *An Investigation of the False Discovery Rate and the Misinterpretation of* P *Values*, 1 R. SOC. OPEN SCI. 1, 11 (2014).

[77] Ronald A. Fisher, *The Arrangement of Field Experiments*, 33 J. MIN. AGRIC. G. BR. 503, 504 (1926).

proper conclusions from individual studies.[78]

As should be obvious, it is the contention here that what Goodman describes is exactly the situation that now exists in the legal system.

It is by now widely recognized that what passes for "findings" in science are not "established" in the way they are often thought to be. In a provocative essay titled *Why Most Published Research Findings Are False*, John Ioannidis noted the increasing concern in modern research that "false findings may be the majority or even the vast majority of published research claims."[79] One reason is the claiming of conclusive research findings on "the basis of a single study assessed by formal statistical significance, typically for a *p*-value of less than 0.05"[80]—in other words, just the kind of study that often forms the centerpiece of an employment discrimination plaintiff's case (although typically with substantially more attention to potentially moderating variables than exhibited in the typical discrimination case). David Colquhoun estimates that 30% is the minimum bound of the proportion of experiments wrongly claiming to have an effect and that Ioannidis's assertion that a majority of published research findings is false "seems to be not unduly alarmist."[81] One reason is that it is relatively easy to finesse an analysis to convert not-quite-significant results into statistical significance.[82]

Far from establishing anything to any degree of certainty, statistically significant results in the social-science literature are increasingly recognized as being of questionable merit. As one statistician has noted, "The ease with which a researcher can report statistically significant evidence for untrue hypotheses fills the literature with false positives."[83] Recent attempts to replicate results of published studies—even well-regarded ones—have often failed.[84] An attempt at replication of 100 studies published in three leading

---

[78] Goodman, *supra* note 24, at 998.

[79] John P.A. Ioannidis, *Why Most Published Research Findings Are False*, 2 PLOS MED. 696, 696 (2005).

[80] *Id.*; *see also* Regina Nuzzo, *Statistical Errors: P Values, the 'Gold Standard' of Statistical Validity, Are Not as Reliable as Many Scientists Assume*, 506 NATURE 150, 150 (2014) (referring to "the surprisingly slippery nature of the *P* value, which is neither as reliable nor as objective as most scientists assume"); Lewis G. Halsey et. al., *The Fickle P Value Generates Irreproducible Results*, 12 NATURE METHODS 179, 179 (2015).

[81] Colquhoun, *supra* note 76, at 13.

[82] *See* E. J. Masicampo & Daniel R. Lalande, *A Peculiar Prevalence of P Values Just Below .05*, Q. J. EXP. PSYCHOL. 2271, 2272 (2012), http://dx.doi.org/10.1080/17470218.2012.711335; Joseph P. Simmons et. al., *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant*, 22 PSYCHOL. SCI. 1359, 1359 (2011), http://journals.sagepub.com/doi/pdf/10.1177/0956797611417632; Leslie K. John et. al., *Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling*, 23 PSYCHOL. SCI. 524, 524 (2012), http://journals.sagepub.com/doi/pdf/10.1177/0956797611430953.

[83] KLINE, *supra* note 14, at 106; *see also* Simmons et. al., *supra* note 82, at 1359 (noting that "it is unacceptably easy to publish 'statistically significant' evidence consistent with any hypothesis").

[84] Harold Pashler & Eric-Jan Wagenmakers, *Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?*, 7 PERSP. ON PSYCHOL. SCI. 528, 528 (2012) (describing current concern over reliability of research findings in psychology), http://www.ejwagenmakers.com/2012/PashlerWagenmakers2012.pdf.

psychology journals found that only a minority could be replicated with statistically significant results.[85]

Given the relatively low standards applied to social science findings of statistically significant differences, it is disappointing to see the eagerness of courts to accept evidence that would not even be sufficient to a social scientist. The Supreme Court in *Bazemore v. Friday* encouraged such reasoning when it observed, in rejecting a defense argument about the weakness of the plaintiffs' statistical showing, that a "[a] plaintiff in a Title VII suit need not prove discrimination with scientific certainty; rather, his or her burden is to prove discrimination by a preponderance of the evidence."[86] Although literally true—though nothing in science is "certain" and amenable to "proof" the way a mathematical theorem is—the Court's statement seemed to be animated by an overly sanguine view of the certainty provided by scientific studies. The Tenth Circuit has expressed a similar view: "while social scientists search for certainty, the trier of fact in a Title VII case need only find that discrimination is more likely than not."[87] It also noted that "statistics that are insignificant to the social scientist may well be relevant to a court."[88] The underlying theme of these comments is that statistically significant findings in the sciences and social sciences are in some sense "scientifically proved" or "scientifically certain" and that the law requires something less.[89]

Judge Posner has also contrasted science versus law in responding to a disparity that was not significant at the .05 level. In *Kadas v. MCI Systemhouse Corp.*,[90] he stated:

> The 5 percent test is arbitrary; it is influenced by the fact that scholarly publishers have limited space and don't want to clog up their journals and books with statistical findings that have a substantial probability of being a product of chance rather than of some interesting underlying relation between the variables of concern. Litigation generally is not fussy about evidence; much eyewitness and other nonquantitative evidence is subject to significant possibility of error, yet no

---

[85] OPEN SCI. COLLABORATION, *Estimating the Reproducibility of Psychological Science*, 349 SCIENCE 943, 943 (2015).

[86] 478 U.S. 385, 400 (1986).

[87] Pitre v. W. Elec. Co., Inc., 843 F.2d 1262, 1269 (10th Cir. 1988).

[88] *Id.*

[89] *See also* EEOC v. American Nat'l Bank, 652 F.2d 1176, 1192 (4th Cir. 1981) (noting that "authority can be found for the proposition that most social scientists, applying laboratory rigor to rule out chance as even a theoretical possibility rather than the law's rougher gauge of the 'preponderance of the evidence,' are prepared to discard chance as an hypothesis when its probability level is no more than 5%, *i.e.* at approximately two standard deviations"); Marcel C. Garaud, *Legal Standards and Statistical Proof in Title VII Litigation: In Search of a Coherent Disparate Impact Model*, U. PENN. L. REV. 455, 456 (1990) (complaining that disparate-impact plaintiffs have "had to present statistical analyses that satisfy a strict scientific standard that is inconsistent with the traditional legal standards of proof").

[90] 255 F.3d 359 (7th Cir. 2001).

> effort is made to exclude it if it doesn't satisfy some counterpart to the 5 percent significance test.[91]

In fact, however, concerns about reliability of statistical evidence *should* lead to exclusion if the statistical study does not satisfy the dictates of Federal Rule of Evidence 702, just as with other scientific or technical evidence. Rule 702 actually *requires* judges to be "fussy" about evidence. As the Supreme Court stated in *Kumho Tire Co. v. Carmichael*,[92] the purpose of the judge's gatekeeping role "is to make certain that an expert, whether basing testimony upon professional studies or personal experience, employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field."[93] If a disparity that does not provide a *p*-value of less than .05 would not be accepted as meaningful in the expert's discipline, it is not clear that the expert should be allowed to testify—on the basis of his expertise in that discipline—that the disparity is, in fact, meaningful.[94]

Although one might question why scientific evidence should have to meet a higher standard than other evidence, the heightened reliability standard for expert testimony established by the Supreme Court's decision in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*[95] and revised Rule 702 is justified on the ground that "expert testimony is . . . uniquely vulnerable to 'adversarial bias'"[96]—that is, the fact that reports are produced by experts retained by the opposing parties who have financial incentives to tailor their testimony in a particular direction. Adversarial bias is a particular risk with statistical analyses, given the flexibility that analysts have in interpreting the results of studies that they themselves have run, as opposed to reporting on studies performed and interpreted by others.[97] There is a broader kind of bias, though, and that is in the research literature itself. As John Ioannidis has noted, bias may not have financial roots, but rather "[s]cientists in a given field may be prejudiced purely because of their belief in a scientific theory or commitment

---

[91] *Id.* at 362. Elsewhere, Judge Posner was more explicit on this point, arguing that "excluding statistical evidence that failed to reach the five percent significance level would imply that eyewitness testimony, too, should be inadmissible unless the probability that the testimony would have been given even if the event testified to had not occurred was less than five percent"). Posner, *supra* note 56, at 1511. The practiced eye will now see that Judge Posner was engaging in the transposition fallacy.

[92] 526 U.S. 137, 152 (1999).

[93] *Id.*; *see also* FED. R. EVID. 702 advisory committee's note to 2000 amendments (embodying the direction of the *Kumho* Court's holding in regards to expert testimony).

[94] *See* David L. Faigman et. al., *Group to Individual (G2i) Inference in Scientific Expert Testimony*, 81 U. CHI. L. REV. 417, 464 (2014) (noting that expert's "[a]ssertion that 'the journal publication context' requires 'a higher standard of proof' than the courtroom starkly indicates a failure to apply in the courtroom 'the same level of intellectual rigor that characterizes the practice of an expert in the relevant field'") (quoting *Kumho Tire*, 526 U.S. at 152).

[95] 509 U.S. 579, 595 (1993).

[96] *Id.*; *see* David E. Bernstein, *Expert Witnesses, Adversarial Bias, and the (Partial) Failure of the* Daubert *Revolution*, 93 IOWA L. REV. 451 (2008).

[97] *See generally* Simmons et. al., *supra* note 82, at 1359, http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124.

to their own findings."[98] This kind of bias is very likely to color the kinds of research findings about discrimination that are published.[99] It is probably fair to say that most social scientists who study discrimination are invested in there being a lot of it for reasons both ideological (it suits their world-view)[100] and pragmatic (findings of no or low levels of discrimination are far less sexy, and thus less publishable, than findings of rampant discrimination, a fact that creates pressure on researchers to resort to increasingly broad definitions of discrimination in order to obtain "significant" results).[101] As David Hull has observed, "One of the strengths of science is that it does not require that scientists are unbiased, only that different scientists have different biases."[102] Unfortunately, in many of the social sciences, most scientists have the same bias.[103]

Statistical studies introduced at trial by experts in their fields often fall far short of the standards of publishable work even in the social sciences. For example, in *Dobbs-Weinstein v. Vanderbilt University*,[104] the plaintiff's expert, a well-known sociologist, submitted a report concluding that her regression analysis demonstrated sex discrimination in setting salaries of university faculty. Yet, according to the District Court, she omitted from her analysis "most of the measures of professional achievement—such as research productivity, speaking engagements, participation in professional organizations and other service to the

---

[98] Ioannidis, *supra* note 79, at 698.

[99] In her article, Deborah Weiss discusses the work of William Bielby, whom she describes as "both a leading scholar on discrimination and a leading expert for plaintiffs." Whether one thinks that ideological bias or financial bias is a greater risk to the integrity of analyses, Bielby embodies both. He has probably earned hundreds of thousands, if not millions, of dollars testifying for plaintiffs in discrimination cases, and he relies to a large extent on his work that is published in scholarly journals. If his scholarly work took a turn less favorable to plaintiffs, one might imagine that this abundant source of cash would dry up. Weiss, *supra* note 28, at 1684.

[100] José Duarte et. al., *Political Diversity Will Improve Social Psychological Science*, 38 BEHAV. & BRAIN SCI. 1, 1 (2015) (describing the lack of political diversity in social psychology and its impact not only on the interpretation of results but also on the nature of the questions that are asked).

[101] *Cf.* Green, supra note 56, at 92 ("[R]egulation of some of the more complex, subtle forms of discrimination common in today's workplace requires a focus on the operation of discriminatory bias as influenced, enabled, and even encouraged by the structures, practices, and dynamics of the organizations and groups within which individuals work. In other words, it posits the need to conceptualize discrimination in terms of *workplace dynamics* rather than solely in existing terms of an identifiable actor's isolated state of mind, a victim's perception of his or her work environment, or the job-relatedness of a neutral employment practice with adverse consequences." (footnote omitted)); *see also* Samuel R. Bagenstos, *Implicit Bias, "Science," and Antidiscrimination Law*, 1 HARV. L. & PUB. POL. REV. 470, 478 (2007) (arguing that "responding to implicit bias requires moving 'beyond the generally accepted normative underpinnings of antidiscrimination law'") (citation omitted); Jerry Kang & Kristin Lane, *Seeing Through Colorblindness: Implicit Bias and the Law*, 58 UCLA L. Rev. 465 (2010).

[102] DAVID L. HULL, SCIENCE AS A PROCESS: AN EVOLUTIONARY ACCOUNT OF THE SOCIAL AND CONCEPTUAL DEVELOPMENT OF SCIENCE 22 (2010).

[103] *See generally* Duarte et al., *supra* note 100. An interesting phenomenon in the scholarly literature is that the "harder" the science, the more likely researchers are to report negative results, or conversely, that the "softer" the discipline the more likely researchers are to report positive results. Daniele Fanelli, *'Positive' Results Increase Down the Hierarchy of the Sciences*, 5 PLOS ONE 5(4), 2 (2010). This result might seem counter-intuitive, since researchers in harder sciences might be expected to formulate more theoretically rigorous hypotheses. Yet, researchers in softer fields have more freedom to interpret their results, thereby giving them the opportunity to "find" in the data results they believe true.

[104] 1 F. Supp. 2d 783 (M.D. Tenn. 1998).

field, and success in obtaining grants—which contribute to disparities in salary."[105] Moreover, most of the disparities that the expert identified were not statistically significant, which was the basis for the District Court's discounting of her analysis.[106] One assumes (or at least hopes) that sociology journals would demand more from a scholar who was purporting to demonstrate the existence of discrimination.[107]

One difference between what courts do and what professional standards in the social sciences require is how they respond to statistical analyses that do not really prove much. Although social-science journals do not reject every article that lacks a compelling statistical analysis, they will reject them when the data upon which the analysis is based are so incomplete that no meaningful conclusions can be drawn from them. In contrast, courts often take the attitude that if the only data are incomplete (especially if it is through no fault of the plaintiff), it would be unfair not to let the plaintiff rely on even bad data, leaving it to juries to sort out the inadequacies of the studies, as if they were capable of doing so.[108] If a social scientist submitted a study to a journal with a small sample showing no statistically significant results, for example, it would probably be returned to the investigator, along with suggestions that the sample size needed to be increased in order to provide meaningful results. In contrast, courts often forgive non-significant results in plaintiff's regressions on the ground that the sample is so small,[109] when they

---

[105] *Id.* at 805; *see also* Rudebusch v. Hughes, 313 F.3d 506, 528 (9th Cir. 2002) (Klinefeld, J., concurring in part and dissenting in part) (noting that "[a] regression of salaries in a university setting that doesn't include doctorate, merit, or performance as variables is like a regression study that predicts shoe size from weight without considering foot size").

[106] Despite the court's acknowledgment of the study's inadequacies, it suggested that its conclusions about the study would have been "entirely different" had the disparities been statistically significant, as if statistical significance could make up for the failure of a model to reflect reality. *Dobbs-Weinstein*, 1 F. Supp. 2d at 807 n.34.

[107] *See also* Randall v. Rolls-Royce Corp., 637 F. 3d 818, 822 (7th Cir. 2011) (plaintiff's expert testified that employer's compensation system was discriminatory without controlling for differences in jobs performed by male and female employees in each compensation category; when those were controlled, the sex disparity disappeared).

[108] *See, e.g.*, Malave v. Potter, 320 F.3d 321, 323 (2d Cir. 2003) (where plaintiffs in a promotion case did not have data on the applicant pool or the eligible labor pool, Court of Appeals overturned summary judgment for defendant on the ground that comparisons to the ethnic composition of the overall Postal Service workforce in Connecticut were good enough); EEOC v. Morgan Stanley & Co., 324 F. Supp. 2d 451, 458 (S.D.N.Y., 2004) (rejecting defendant's argument that the EEOC's expert's report should be excluded for failure to include variables such as the employee's ability to interact productively with other employees; assistance provided to co-workers; work ethic; participation in firm recruiting, mentoring and training; and competitive market dynamics. The court reasoned that because no data existed on these variables that the employer contended were important in promotion and compensation, a regression that did not consider them should be permitted to go to the jury, leaving the jury to decide which variables should be included).

[109] *See, e.g.*, Chin v. Port Auth. of N.Y. & N.J., 685 F.3d 135, 153 (2nd Cir. 2012) (accepting a study with a *p*-value of .13, stating that "requiring a statistical showing of 95-percent confidence would make it mathematically impossible to rely upon statistics in a case like this one, in which the relevant population included so few Asian Americans"); Waisome v. Port Auth., 948 F.2d 1370, 1379 (2d Cir.1991) (noting that "the lack of statistical significance in the ultimate promotion reflects only the small sample size."). *See also* Transcript of Videotaped Deposition of Michael A. Campion, Ph.D., at 184, Karlo V. Pittsburgh Glass Works, LLC, Case 2:10-cv-01283-TFM Document 317-4, W.D. Pa., Aug. 2, 2013, 2014 WL 12539666 (testifying that results are significant at the thirteen percent level).

should recognize that some questions—especially when sample size is small—simply are not amenable to statistical analysis. The problem with courts' not fulfilling their gatekeeping function and allowing juries to decide on their own is that, as Judge Jack Weinstein has stated, "An expert can be found to testify to the truth of almost any factual theory, no matter how frivolous . . . ."[110] Judges who fail in their gatekeeping responsibility then simply throw up their hands and call it a "jury question."

Judge Posner's relaxed attitude toward issues of statistical significance in *Kadas* was recently taken to an absurd extreme by the District Court in *In re Photochromic Lens Antitrust Litigation*. Judge Posner had said in *Kadas*, after noting that the five-percent significance level was arbitrary—which it surely is—that "[i]t is for the judge to say, on the basis of the evidence of a trained statistician, whether a particular significance level . . . is too low to make the study worth the consideration of judge or jury."[111] In the *Photochromic Lens* case, the magistrate judge had excluded an expert's studies because he had used a *p*-value of .50 – yes, *.50*! The district court rejected the magistrate's conclusion, accepting the expert's testimony that he had used a significance level of .50 for two reasons: 1) because the data set was so limited that using a more conventional significance level would mean that the study would lack the "power to detect impact"; and 2) because he wanted to avoid false negatives, *i.e.*, Type II errors.[112] It should first be noted that these are not two different reasons, but rather two ways of stating the same reason because power is the likelihood of not making a Type II error.[113] Moreover, given that it is axiomatic that there is a tradeoff between false positives (Type I errors) and false negatives (Type II errors) (although the extent of the tradeoff is generally impossible to quantify), a desire to avoid Type II errors would *always* counsel in favor of a more permissive significance level.

No competent social scientist would use a significance level of .50 (or anything close to it, for that matter) because the null hypothesis would be rejected in too many cases in which it is true—with a significance level of .50, in almost half of all such cases.[114] To return to the coin-flip analogy, imagine that you wanted to test whether a coin was fair. You flip the coin

---

[110] Jack B. Weinstein, *Improving Expert Testimony*, 20 U. RICH. L. REV. 473, 482 (1986); *see also* KLINE, *supra* note 14, at 100 (observing that "[t]he sad truth is that there is no claim so preposterous that a PhD scientist cannot be found to vouch for it") (quoting Robert L. Park, *The Seven Warning Signs of Voodoo Science*, 1 THINK 33, 33 (2003)).

[111] Kadas v. MCI Systemhouse Corp., 255 F.3d 359, 363 (7th Cir. 2001).

[112] In re Photochromic Lens Antitrust Litigation, (M.D. Florida April 3, 2014) WL 1338605; 2014-1, Trade Cases P 78, 732.

[113] Jacob Cohen, *A Power Primer*, 112 PSYCHOL. BULL. 155, 156 (1992). The probability of making a Type II error–that is, the probability of failing to reject a false null hypothesis–is denominated β. Power is defined as $1 - \beta$. *See id.*

[114] *But see* Richard Lempert, *Statistics in the Courtroom: Building on Rubinfeld*, 85 COLUM. L. REV. 1098, 1099 (1985) (asserting that "[s]urely statistical evidence that is significant at the .10 level or even the .50 level often meets" the relevance test of Rule 401).

100 times. In order to reject the null hypothesis using a significance level of .05, you must get a result more extreme than a 60-40 split. In order to reject the null hypothesis at a .50 level, you need only a result that is more extreme than a 53-47 split.[115] No one who knows the slightest amount of statistics would actually believe that one should label as "statistically significant" a disparity that would be expected by chance about half the time, although perhaps such a person would be willing to so testify if the price is right.[116] It is true, as the experts claimed, that the risk of false exculpations is greatly reduced by setting the significance level so high—but the probability of *any* exculpation at all, even a correct one, is also dramatically reduced.

The comparison between scientific and legal proof is based on both an overestimation of the certainty of what statistically significant results mean in science and the meaningfulness of statistical analyses in legal cases. The notion that the law should be more lax than the relatively lax standards of social science leads to admission into evidence of statistical analyses that do not mean very much. A social science study that reports a spurious finding wastes a few pages in a journal that were likely to be wasted anyway. Yet, millions of dollars often turn on whether a particular statistical study is accepted as establishing a prima facie case of discrimination.[117] The Ninth Circuit (oddly enough) had it right in *Penk v. Oregon State Board of Higher Education*,[118] stating:

> We note in passing that it is often acceptable in the social sciences to use a statistical model for proof of behavior even where, in an absolute sense, that model does not describe the data well. However, courts are not free to decide legal propositions on hypothetical evidence. This is especially true

[115] This is using a two-tailed test. Using a one-tailed test, the results are even starker. Using a one-tailed test, you are testing for disparity in only one direction. So, assume that you are testing whether the coin is biased toward heads. Because you want a really good test, you give it a million trials (you have a lot of time and a really strong thumb). You toss 500,001 heads out of 1,000,000 trials, a result far closer to the expected 500,000 than one could ever realistically hope to obtain in the real world. Yet, the disparity between observed and expected would be "statistically significant" at the .50 level—that is, the result would be obtained by chance less than 50 percent of the time assuming that the null hypothesis (that the coin was fair) is true.

[116] One might have thought that there was only one expert in the world willing to testify that a *p*-value of .50 was appropriate. One would have been wrong. In *In re High-Tech Employee Antitrust Litigation*, WL 1351040 (N.D. Cal. 2014), the plaintiffs' expert sought to testify that results of his statistical model were statistically significant at the .50 level and that such a significance level was an appropriate balance of the risk of Type I and Type II errors. The court did not address the merits of the expert's opinion, however, instead excluding it on the ground that the assertion was not raised until a reply brief and was therefore untimely. *Id.* at 7. The court still managed to commit the transposition fallacy along the way, however. *See id.* at 6–7 ("If the p-value is less than or equal to the selected significance level, the null can be rejected because the result is said to be 'statistically significant' at that level, which means the probability that the observed association is the result of chance rather than a true association is less than the stated significance level.").

[117] Either because it leads to an adverse judgment on the merits against the employer or forces settlement. Given the extremely high cost of defending class actions, class certification usually leads to settlement, often resulting in recovery for plaintiffs even in cases they would be unlikely to win at trial. *See* Deborah M. Weiss, *A Grudging Defense of* Wal-Mart v. Dukes, 24 YALE J. L. & FEM. 119, 134 n.56 and accompanying text (2012).

[118] 816 F.2d 458, 468 n.1 (9th Cir. 1987).

where courts are asked to draw inferences as to the existence of hidden discriminatory motives from statistical evidence.

Unlike in law, in science, "behavioural decisions are rarely made on the basis of a single significance test."[119] As Jason Chin has noted, though the scientific stakes of allowing bodies of research to rest on unreliable findings are high, they are just as high in law (one might argue, even higher), where there is less opportunity for self-correction than there is in science.[120]

E. Would a Better Explanation of the Meaning of the P-*Value Solve the Problem?*

If the meaning of the *p*-value that judges and jury seemingly universally rely on is incorrect, one solution might be to make sure that they are correctly educated about its meaning, untainted by the transposition fallacy. It is, after all, the transposition fallacy that causes the false equating of significance levels and standards of proof. It might be argued that except for correcting that small flaw, cases otherwise could proceed much the same as they would today.

There are two problems with such an approach. The first is that it is very difficult to prevent people from engaging in the transposition fallacy, so it is questionable whether in the short space of a trial the jury could be educated out of it. The second is that, once the transposition fallacy is out of bounds, the probative value of *p*-values is quite slim. That is, they simply don't tell you very much.

1. The Transposition Fallacy Seems Intractable

Experience has shown that the hope that jurors can be instructed in such a way as to avoid the transposition fallacy is unrealistic. After all, judges routinely engage in the transposition fallacy despite the fact that it is amply described in the legal literature. The Federal Judicial Center's *Reference Manual on Scientific Evidence*, provided to all federal judges, could not be clearer: "*p* is not the probability of the null hypothesis given extreme data."[121] One employment discrimination casebook in its Seventh Edition[122] noted an earlier article of mine that described the transposition fallacy and observed that I was "theoretically correct," but nonetheless went on to describe the meaning of the *p*-value in terms embodying the transposition fallacy.[123] In the Eighth Edition of

---

[119] Peter Dixon, *The* P-*Value Fallacy and How to Avoid It*, 57 CANAD. J. EXPER. PSYCHOL. 189, 190 (2003).

[120] Jason M. Chin, *Psychological Science's Replicability Crisis and What it Means for Science in the Courtroom*, 20 PSYCHOL., PUB. POL., & LAW, 225, 225 (2014).

[121] KAYE & FREEDMAN, *supra* note 17, at 250.

[122] ZIMMER ET AL., 7th ed., *supra* note 61, at 143–144.

[123] *Id.* at 143 ("Setting the level of significance at 0.05 means that a Type I error is made in only 5 percent of the cases, that is, 5 in 100 times").

the same book, the acknowledgment that I was correct was removed, the transposition fallacy was described in even more cursory fashion, and the meaning of the significance level was described in the same fallacious terms as in the prior edition.[124] It thus appears that even a glimmer of understanding of the issue is easily extinguished in the absence of constant vigilance.

The fallacy persists even in the social sciences among people who clearly should know better. Psychologists Ruma Falk and Charles Greenbaum have suggested why the misconception persists: "When a procedure instructs us to reject a hypothesis, in the context of scientific induction, believing that the hypothesis deserves to be rejected, namely that it is no longer credible, is inevitable."[125] They continue:

> Altogether, misinterpreting a significant result as conveying the probability of the null hypothesis is prevalent and robust in the face of variations in problem presentation and in subjects' level of sophistication. It appears that the illusion is hard to eradicate. It keeps creeping into texts of authors who surely know better, and it is devious in evading critical reviewing. We confess to believing so ourselves for years, and probably passing it over to many of our students, until we came to realize our mistake. Serious intrinsic factors must work to keep the misconception alive.[126]

These comments, it should be noted, refer to the persistence of the transposition fallacy among professionals who perform statistical analyses for a living. If the clear and thorough explanation of *p*-values in the literature has not penetrated the minds of highly educated judges and many social scientists,

---

[124] In the Seventh Edition, there was a description of how the transposition fallacy could result in a number of false positives in discrimination cases far in excess of the five percent that the typical interpretation of the *p*-value would suggest. *Id.* at 144. In the Eighth Edition, the only illustration of how the fallacy operates is with respect to a medical test for a rare disease. ZIMMER ET AL., 8th ed., *supra* note 15, at 135. In the Seventh Edition, the book accurately described the problem (although not calling it the transposition fallacy) as "confus[ing] the probability of a particular result given the null hypothesis with the probability of the null hypothesis given the observed result." ZIMMER ET AL., 7th ed., *supra* note 61, at 144. That explanation was removed from the Eighth Edition, which continues to aggressively promote the transposition fallacy. ZIMMER ET AL., 8th ed., *supra* note 15, at 127 ("Be clear what this means: if a certain result is 'statistically significant,' it is unlikely to be the result of chance."); *id.* at 134 ("The statistician's job is to determine the probability that chance explains the difference"); *id.* ("[A] statistician can inform the court how probable it is that a certain pattern of selections would have occurred if color were not somehow influencing the selection decision."); *id.* at 135 ("Setting the level of significance at 0.05 means that a Type I error is made in only 5 percent of the cases, that is, 5 in 100 times."); *id.* at 137 ("The outcome is not likely to be the result of chance when a result is more than 2 standard deviations from the norm; in such a case, there are only 4 chances in 100 that the result is consistent with the null hypothesis, that the differences are the result of chance."); *id.* at 138 ("Rejecting the null hypothesis means that it is much more likely than not (though not certain) that the null hypothesis is incorrect.").

[125] Ruma Falk & Charles W. Greenbaum, *Significance Tests Die Hard: The Amazing Persistence of a Probabilistic Misconception*, 5 THEORY & PSYCHOL. 75, 81 (1995).

[126] *Id.* at 86; *see* Sellke et al., *supra* note 8, at 62 (noting that "[a]lthough standard textbooks typically warn against such interpretations, the warnings often go unheeded"); *id.* at 71 (noting that "[t]he standard approach in teaching—of stressing the formal definition of a *p* value while warning against its misinterpretation—has simply been an abysmal failure").

the idea that a jury is going to achieve enlightenment on the point during the course of a trial seems purely fanciful.[127]

The persistence of erroneous beliefs about *p*-values was illustrated in a study of the effectiveness of teaching interventions designed to counteract their misunderstanding.[128] The researchers identified two possible reasons for the seeming intractability of the transposition fallacy. The first is simply a lack of understanding of Bayesian posterior probabilities, which leads people to confound the prior probability, which is given by the *p*-value, with the posterior probability. The second is that it is based on a misapplication of the Modus Tollens argument in a probabilistic context. The Modus Tollens argument has the structure:

> If A, then not B,
>
> B
>
> Therefore, not A.

In the context of hypothesis testing, a matching argument would be:

> If the Null Hypothesis is true, the Data would not occur.
>
> The Data have occurred
>
> Therefore, the Null Hypothesis is not true.

But in actual hypothesis testing, the premises are not categorical but probabilistic:

> If the Null Hypothesis is true, the Data are unlikely to occur
>
> The Data have occurred
>
> Therefore the Null Hypothesis is probably not true.

What is a valid argument in the categorical case is not necessarily a valid argument in the probabilistic case, as the researchers point out by relying on the following example:

> If a person is an American, then he is probably not a

---

[127] Gerd Gigerenzer asks:

> Why do intelligent people engage in statistical rituals rather than in statistical thinking? Every person of average intelligence can understand that p(D|H) is not the same as p(H|D). That this insight fades away when it comes to hypothesis testing suggests that the cause is not intellectual but social and emotional.

Gigerenzer, *supra* note 14, at 599. He finds the answer to the question in unconscious conflicts among statistics researchers. But judges and lawyers are not party to those conflicts, and they also seem unable to avoid the trap.

[128] Pawel Kalinowski, Fiona Fidler & Geoff Cumming, *Overcoming the Inverse Probability Fallacy: A Comparison of Two Teaching Interventions*, 4 EXPERIMENTAL PSYCHOL. 152, 152 (2008).

member of Congress

This person is a member of Congress.

Therefore, he is probably not an American.

Given the two potential sources of the transposition fallacy, the researchers designed two interventions to attempt to counter them. Students were exposed to one of two 45-minute tutorials, one explicitly contrasting Bayesian analysis with NHST *p*-values and the other explaining why Modus Tollens reasoning does not apply to probabilistic premises. Subjects were undergraduate students who had completed four courses in statistics and were familiar with *p*-values and NHST. Prior to the intervention, the students were given a survey in which they were asked whether they agreed or disagreed with six typical misinterpretations of *p*-values, and they were then given the survey immediately after the tutorial and then again five weeks later (wording of the questions was altered to attempt to mitigate practice effects).

The interventions did have some effect. Prior to the intervention, the average number of misconceptions about *p*-values was 4.0 per student (out of the maximum of 6), with 97% incorrectly agreeing with at least 1. Immediately after the intervention, the average went down to 2.0 misconceptions, but it went up to 2.7 per student at the five-week follow-up. The researchers concluded that the interventions "were remarkably effective in reducing the misconceptions of NHST *p* values," and it is clear that the post-intervention performance was better than that prior to the intervention.[129]

The glass-half-empty story in the study is that the intervention only reduced the students' misconceptions by half immediately after the intervention and only by a third measured five weeks later, at a time when the average student endorsed almost half (45%) of the tendered misconceptions.[130] Thus, students experienced in statistics who were specifically schooled on the transposition fallacy nonetheless continued to fall into its trap leading to serious doubt that the deep innumeracy of judges and juries could be overcome during the course of a trial.

2. The *P*-Value Does not Tell You Much

Even if the transposition fallacy could be avoided, it is not clear what would be gained by conveying a correct understanding of *p*-values to juries. Informing the trier of fact that the *p*-value is not the probability that an employer obtained its workforce through random selection, but rather the probability that an employer selecting randomly, would have that representation assuming that the null hypothesis is true does not provide the

---

[129] *Id.* at 157.

[130] *See also* KLINE, *supra* note 14, at 104 (noting that "the cliché of 'better teaching' about significance testing has not in more than 60 years improved the situation").

trier of fact with information that is useful. Apart from that definition, however, it is much easier to say what the *p*-value is not than what it is.

The *p*-value is of only modest relevance to someone trained not to engage in the transposition fallacy (though, as has been shown, it is of great relevance to someone who commits the fallacy). As a general matter, it can be said that small *p*-values are in general less consistent with the null hypothesis than large *p*-values. As Geoff Cumming has observed:

> Other things being equal, the smaller the *p*, the more reason we have to doubt the null. In a later chapter I'll demonstrate that *p* is actually an extremely poor and vague measure of evidence against the null. Even though, thinking of *p* as strength of evidence may be the least bad approach.[131]

Now, that approach may lead one to think, for example, that if one employer's workforce has a disparity yielding a *p*-value of .05 and another has a disparity with a workforce yielding a disparity of .01, the null hypothesis is less likely to be true in the latter case than in the former. But that is not correct, since all else is unlikely to be equal.[132] Thus, "the *P* value . . . does not reliably indicate the strength of evidence against the null hypothesis."[133]

Now, it certainly true as a general matter that—given a particular employer—a disparity with a *p*-value of .001 suggests that the null hypothesis is less likely to be true than if the disparity had a *p*-value of .05 (although how much less likely it is cannot be calculated by comparing the two *p*-values). But here, you don't need the lower *p*-value to tell you that the null hypothesis is less likely; you have the more extreme disparity that led to the greater *p*-value to tell you that. The *p*-value adds virtually nothing and just muddies the water.

3. *P*-Values Should Be Excluded under FRE 403

The ease with which reported *p*-values cause a trier of fact to slip into the transposition fallacy and the difficulty of avoiding that lapse of logic, coupled with the relatively sparse information actually provided by the *p*-value, make *p*-values prime candidates for exclusion under Federal Rule of Evidence 403. That rule allows for exclusion of evidence "if its probative value is substantially outweighed by a danger of . . . confusing the issues, [or] misleading the jury . . . ."[134] If judges, not to mention the statistical experts

---

[131] CUMMING, *supra* note 8, at 33.

[132] *See* Steven Goodman, *A Dirty Dozen: Twelve P-Value Misconceptions*, 45 SEMINARS IN HEMATOLOGY 135, 137 (2008) (noting that "[d]ramatically different observed effects can have the same *P* value").

[133] Halsey et al., *supra* note 80, at 179. In order for the *p*-value to give a reliable indication of the strength of evidence against the null hypothesis, the statistical power of the study must be very high. Statistical power is the capacity of the experiment to find an effect when there actually is an effect. The power of a study cannot be measured directly, but it depends on a variety of study features, including the significance level, the size of the expected effect, population variability, the nature of the alternative hypothesis, the nature of the test, and sample size. *Id.* at 181.

[134] FED. R. EVID. 403.

they rely on, cannot use the information without falling into fallacious reasoning, the likelihood that the jury will misunderstand the evidence is very high. Since the *p*-value actually provides little useful relevant information, the high risk of misleading the jury greatly exceeds its scant probative value, so it simply should not be presented to the jury.[135]

One imagines that plaintiffs' lawyers will complain mightily about exclusion of *p*-values because they have proven so helpful to them. However, the attractiveness of *p*-values in litigation derives specifically from the transposition fallacy. The fact-finder is trying to decide whether the employer's workforce distribution was caused by chance or something else and to make the decision using a preponderance-of-the-evidence standard, which is generally defined as more likely than not (50+% chance). To learn that the probability of "something" is less than five percent bears meaningfully on the question before the fact-finder *only* if it takes that "something" to be the likelihood of random selection or the likelihood that the employer did not discriminate.

As a case in point, consider the plaintiff's expert's testimony in *Tabor v. Hilti, Inc.*[136] The employer's statistical disparity was 2.777 standard deviations from the "expected," and the expert testified (not engaging in the transposition fallacy) that "the probability is less than 0.006 that a disparity at least as large as this could occur solely as the result of chance factors, if promotions were unrelated to sex."[137] What could the jury make of that 0.006 probability, if not the probability that chance was responsible? Jurors know that the evidence presented to them has been screened for relevance, so their minds are primed to attach significance to it.[138]

It is probably not overstating the case to say that there is simply nothing a jury can do with the *p*-value other than misuse it or ignore it. The best course is to force them to ignore it by excluding it from evidence.

III. BEYOND THE *P*-VALUE: OTHER PROBLEMS WITH NULL HYPOTHESIS SIGNIFICANCE TESTING

Although the transposition fallacy may be the greatest problem with NHST, it is not the only one. In fact, while use of hypothesis testing, *p*-values, and significance levels proceeds apace in the legal world, there has been

---

[135] *See* Daubert v. Merrell Dow Pharma., Inc., 509 U.S. 579, 595 (1993) ("Expert evidence can be both powerful and quite misleading because of the difficulty in evaluating it. Because of this risk, the judge in weighing possible prejudice against probative force under Rule 403 of the present rules exercises more control over experts than over lay witnesses.") (quoting Jack B. Weinstein, *Rule 702 of the Federal Rules of Evidence is Sound; It Should Not Be Amended*, 138 F.R.D. 631, 632 (1991).

[136] *See* 703 F.3d 1206, 1223 (10th Cir. 2013).

[137] *Id.* at 1223. Although the expert's testimony did not incorporate the transposition fallacy, the court's description of the meaning of statistical significance did. *Id.* ("Statistical significance measures the likelihood that the disparity between groups is random, i.e., solely the result of chance.").

[138] Pun intended.

substantial reconsideration of their use in the sciences and social sciences, from which the law initially derived their use.[139]

Responding to increasing concerns over NHST, the journal *Basic and Applied Social Psychology* recently announced that it would no longer publish papers using NHST, along with its vestiges, including *p*-values and the labeling of differences as "statistically significant."[140] Invoking the transposition fallacy, although not calling it that, the editors of *Basic and Applied Social Psychology* reasoned that the primary problem with NHST is the problem "traversing the distance from the probability of the finding, given the null hypothesis, to the probability of the null hypothesis, given the finding."[141] Responding to the concern that, with the ban on NHST, it might be easier to publish in the journal, they argued to the contrary: "[W]e believe that the $p<.05$ bar is too easy to pass and sometimes serves as an excuse for lower quality research."[142] Thus, the editors suggest supplying descriptive statistics but not inferential ones.[143]

Among the problems identified with NHST, in addition to the transposition fallacy (which everyone who has thought seriously about it understands is a problem), include the fact that the null hypothesis is often quite unlikely to begin with, that NHST tends to encourage conflation of statistical and practical significance, that deriving any meaning from the *p*-value at all requires that the model be accurately specified, and that the alternative hypothesis sought to be proved is often not the only alternative that may follow from rejection of the null hypothesis.

---

[139] *See* CUMMING, *supra* note 8; KLINE, *supra* note 14; Goodman, *supra* note 24; Nuzzo, *supra* note 21. *But see* Richard L. Hagen, *In Praise of the Null Hypothesis Statistical Test*, 52 AM. PSYCHOL. 15, 22 (1997) (defending the continued use of NHST). Hagen acknowledges that "the NHST has been misinterpreted and misused for decades," but argues, "[t]his is our fault, not the fault of the NHST." *See id.* It's important to note that even defenders of using *p*-values do not defend the transposition fallacy. *See* Clarice R. Weinberg *It's Time to Rehabilitate the P-Value*, 12 EPIDEMIOLOGY 288, 288 (2001) ("[The *p*-value] is *not* the probability that a null hypothesis is true, nor is it the probability that we are making a certain error.").

[140] Chris Woolston, *Psychology Journal Bans P Values*, 519 NATURE 9, 9 (March 9, 2015). http://www.nature.com/polopoly_fs/1.17001!/menu/main/topColumns/topLeftColumn/pdf/519009f.pdf. Ironically, even this announcement in *Nature* fell afoul of the transposition fallacy. The following "clarification" was appended to its original story:

> Clarified: This story originally asserted that "The closer to zero the *P* value gets, the greater the chance the null hypothesis is false." *P* values do not give the probability that a null hypothesis is false, they give the probability of obtaining data at least as extreme as those observed, if the null hypothesis was true. It is by convention that smaller *P* values are interpreted as stronger evidence that the null hypothesis is false. The text has been changed to reflect this.

[141] David Trafimow & Michael Marks, 37 BASIC & APPLIED SOC. PSYCHOL. 1, 1 (2015).

[142] *Id.* at 1–2.

[143] *See also* Eric L. Eich, *Business Not as Usual*, 25 PSYCHOL. SCI. 3 (2014) (urging researchers to shift away from their reliance on NHST); *Instructions for Authors*, EPIDEMIOLOGY, http://edmgr.ovid.com/epid/accounts/ifauth.htm ("For estimates of causal effects, we strongly discourage the use of categorized *P*-values and language referring to statistical significance . . . .").

A. The Null Hypothesis Is Unlikely a Priori.

One of the common objections to NHST is that it is "a trivial exercise" because the null hypothesis is always false at some decimal place.[144] The null hypothesis assumes that there is *no* difference between two groups, and there is virtually always *some* difference.[145] David Bakan recounted an analysis of results on a battery of tests given to a nationwide sample of 60,000 subjects:

> Every test came out significant. Dividing the cards by such arbitrary criteria as east versus west of the Mississippi River, Maine versus the rest of the country, North versus South, etc., all produced significant differences in means. In some instances, the differences in the sample means were quite small, but nonetheless, the p values were all very low.

As he noted, "*there is really no good reason to expect the null hypothesis to be true in any population.*"[146] Thus, "when a null hypothesis offers an implausible description of reality, rejecting it provides no information."[147]

In the discrimination context, the null hypothesis is that members of different groups—blacks vs. whites, males vs. females, older vs. younger, etc.—have exactly the same probability of selection in the absence of discrimination, so that a finding of different selection rates is prima facie evidence of invidious discrimination.[148] Given race[149] and sex[150] differences

---

[144] Roger E. Kirk, *Practical Significance: A Concept Whose Time Has Come*, 56 EDUC. & PSYCHOL. MEAS. 746, 747 (1996), http://journals.sagepub.com/doi/pdf/10.1177/0013164496056 005002; *see also* John W. Tukey, *The Philosophy of Multiple Comparisons*, 6 STAT. SCI.100, 100 (1991).

[145] David Bakan, *The Test of Significance in Psychological Research*, 66 PSYCHOL. BULL. 423 (1966), https://projecteuclid.org/download/pdf_1/euclid.ss/1177011945.

[146] *Id.* at 426 (emphasis in original).

[147] Andreas Schwab et al., *Researchers Should Make Thoughtful Assessments Instead of Null-Hypothesis Significance Tests*, 22 ORGANIZATION SCI. 1105, 1107 (2011), http://psycnet.apa.org/journals/bul/66/6/423.pdf.

[148] *See* Adams v. Ameritech Serv., Inc., 231 F.3d 414, 424 (7th Cir. 2000):

> Studies in employment discrimination cases typically begin by defining the relevant labor market, and then ask what the results would be for the salient variable (race in *Mister*, age in our case) if there were no discrimination. That is called the "null hypothesis." If the relevant market is 40% African-American, for instance, one would expect 40% of hires to be African-American under the null hypothesis.

[149] *See, e.g.*, Richard E. Nisbett et al., *Intelligence: New Findings and Theoretical Developments*, 130 AM. PSYCHOL. 130, 130 (2012) (emphasizing social factors as explanations for well-known racial differences in cognitive performance); J. Philippe Rushton & Arthur R. Jensen, *Thirty Years of Research on Race Differences in Cognitive Ability*, 11 PSYCHOL., PUB. POL. & LAW 235, 235 (2005) (emphasizing genetic factors as explanations for well-known racial differences in cognitive performance). Although some scholars have suggested that racial differences in test results reflect test bias, *see, e.g.*, Helen A. Moore, *Whiteness: Some Critical Perspectives: Testing Whiteness: No Child or No School Left Behind?*, 18 WASH. U. J.L. & POL'Y 173 (2005); Susan Sturm & Lani Guiner, *The Future of Affirmative Action: Reclaiming the Innovative Ideal*, 84 CAL. L. REV. 953 (1996), empirical evidence is unkind to that contention, *see* Paul R. Sackett, Matthew J. Borneman, and Brian S. Connelly, *High-Stakes Testing in Higher Education and Employment: Appraising the Evidence for Validity and Fairness*, 63 AM. PSYCHOL. 215 (2008). Although the debate over the extent to whether genetic or environmental factors are responsible for race and sex differences in cognitive performance has a number of public-policy implications, for purposes of examining the use of NHST, the important point is that measured race and sex differences in test results cannot help but have implications for occupational outcomes.

[150] *See, e.g.*, Kingsley R. Browne, *Biological Sex Differences in the Workplace: Reports of the End of*

in education, experience, abilities, and interests, we can usually be quite confident that the null hypothesis is not ever *exactly* true—and often it is not even approximately true. With large enough numbers, whatever difference there is will be statistically significant. In the analysis that Bakan described, differences appeared that would not have been predicted *a priori*; in the discrimination context, however, the assumption that various demographic groups would be identical in all ways that are relevant to workplace outcomes is facially implausible. Yet, that is the assumption that leads to a conclusion (or at least a presumption) that the employer has discriminated once chance has been "eliminated" as an explanation by misinterpretation of the *p*-value. But surely it is not true that, once chance has been eliminated, discrimination is the most likely explanation for, say, the sex-ratio in nursing and auto mechanics, or the dearth of black professors of nuclear physics at major research institutions, or, for that matter, the rarity of white males in gender studies.

B. Conflation of Statistical and Practical (or Legal) Significance

An additional problem with NHST is that the significance level is unmoored from any notions of practical significance. A finding is conventionally labeled "significant" if the *p*-value is less than .05, "highly significant" if the *p*-value is less than .01, and sometimes "very highly significant" if the *p*-value is less than .001. But how impressed should we be that results are "significant" or some variation of it? Does statistical significance imply practical significance? Many courts assume that statistical significance implies "legal significance," but should they?

Courts making the leap from statistical significance to practical significance often employ what Geoff Cumming refers to as "the fallacy of the 'slippery slope of significance.'" "An effect is found to be statistically significant, is described, ambiguously, as 'significant,' and then later is

---

Men Are Greatly Exaggerated (As are claims of Women's Continued Inequality), 93 B.U.L.R. 769 (2013) [hereinafter Browne, *End of Men*] (discussing sex differences in interest and aptitudes that affect employment outcomes); Kingsley R. Browne, *Evolutionary Psychology and Sex Differences in Workplace Patterns, in* GAD SAAD, EVOLUTIONARY PSYCHOLOGY IN THE BUSINESS SCIENCES, 71 (2011) [hereinafter Browne, *Workplace Patterns*] (same); Kingsley R. Browne, *Evolved Sex Differences and Occupational Segregation*, 27 J. ORG. BEHAV. 143 (2006) (same); Kingsley R. Browne, *Women in Science: Biological Factors Should Not Be Ignored*, CARDOZO WOMEN'S L.J., 11(3):509–28 (2005) (same); Kingsley R. Browne, *Biology at Work: Rethinking Sexual Equality* (2002) (same); Kingsley R. Browne, *Sex and Temperament in Modern Society: A Darwinian View of the "Glass Ceiling" and the "Gender Gap" in Compensation*, 37 ARIZ. L. REV. 971 (1995) (same). The existence and effect of differences attributable to sex in the workplace are controversial issues that frequently generate emotional reactions. *See, e.g.*, James Damore, *Why I Was Fired by Google*, WALL ST. J., Aug. 12, 2017, at C2 (employment at Google terminated after the author's circulation within the workplace of a document suggesting that some male-female disparity in technical fields could be attributed to biological differences). Scientists acknowledge that there is significant debate about the origins, meaning, magnitude, and effect of sex differences – but not about the existence of sex differences. *See, e.g.*, Larry Cahill, *An Issue Whose Time Has Come*, 95 J. NEUROSCIENCE RESEARCH 12 (2017); Daphna Joel and Margaret M. McCarthy, *Incoporating Sex as a Biological Variable in Neuropsychiatric Research: Where Are We Now and Where Should We Be?*, 42 NEUROPSYCHOPHARMACOLOGY 379 (2017).

discussed as if it had thereby been shown to be 'important' or 'large.'"[151] But sometimes the proper response to a showing that a difference is statistically significant is "so what"?

The leap from statistical significance to practical or legal significance should not be automatic. One problem is that with large samples, even trivial differences can be statistically significant. Assume an employer with 10,000 employees that hires from a pool that is 50 percent male and 50 percent female. We would "expect" the employer to have 5,000 males and 5,000 females. But suppose the employer has 5,100 males and 4,900 females? Should we take that disparity to establish a prima facie case of a pattern or practice of discrimination? Although it seems unlikely that an employer with such a workforce could be said to have a "standard operating procedure" of discrimination against women, the disparity is statistically significant at the .05 level and for that reason alone would be accepted by many courts as establishing a prima facie case.[152] Yet, if one believes that the proper sex composition would be 5,000 of each sex, that would mean that the employer made 10,000 hiring decisions out of which 100 were discriminatory. It is hard to see how something that the employer does one or two percent of the time can be said to be its standard operating procedure. The effect of large numbers on statistical significance has led Rick Jacobs and colleagues to suggest that liability for employment discrimination may turn more on the fact that the defendant is large than that the defendant is actually discriminating.[153]

Small *p*-values derived from large numbers are not inherently meaningful. In *Bew v. City of Chicago*, a disparate-impact case, the District Court held that the plaintiffs had established a prima facie case of disparate impact where the pass rate on the test for minorities was 98.60 percent compared to the pass rate for whites of 99.95 percent. Most courts would probably have said that a disparate impact had not been demonstrated, because the pass rate for minorities was almost 99 percent that of whites.[154] Yet, the *Bew* court found a disparate impact because the disparity in *fail* rates was greater than 5 standard deviations (corresponding to a *p*-value of less than

---

[151] CUMMING, *supra* note 8, at 29.

[152] *Cf.* Apsley v. Boeing Co., 691 F.3d 1184, 1200–01 (10th Cir. 2012) (although rejecting the argument that there is a "practical significance" requirement, holding that where employer had recommended and hired 99% of the "expected" number of older employees, a jury could not find that discrimination was the company's standard operating procedure).

[153] Rick Jacobs, Kevin Murphy & Jay Silva, *Unintended Consequences of EEO Enforcement Policies: Being Big Is Worse than Being Bad*, 28 J. BUS. PSYCHOL. 467, 469 (2013).

[154] Under the EEOC's "4/5ths Rule," a disparate-impact is generally not shown unless the pass rate for minorities is less than 4/5ths (80%) that of the highest-scoring group, although smaller differences "may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms." 29 C.F.R. § 1607.4(D) (2016). In Jones v. City of Boston, 752 F.3d 38, 43 (1st Cir. 2014), the First Circuit rejected the argument that there is any "practical significance" requirement under the disparate-impact test, thereby holding that a disparate impact could be shown by a relatively trivial difference–6 of 529 black officers and cadets (1.1%) tested positive for drugs, while 3 of 1260 white officers and cadets (0.2%) in a given year tested positive.

1 in a million). It is hard to see how the test in question constituted a substantial barrier to minority achievement of the kind that the Court disapproved in *Griggs*[155] and its progeny. Yet, statistical significance was extremely high, because the fail rate of blacks was about 28 times that of whites. Ironically, this imbalance in fail rates was probably a consequence of the employer's attempt to shrink the racial difference in pass rates by making the test easier (to such an extent that almost everyone passed the test, thereby almost completely undermining the purpose of the test), because shrinking the difference in pass rates will virtually always result in a greater imbalance in fail rates.[156]

C. *Interpretation of the* P-*Value as a Probability Assumes that the Model is Perfectly Specified*

Even an interpretation of the *p*-value that does not embody the transposition fallacy nonetheless is incomplete in its statement that the *p*-value represents the probability of a result as extreme as obtained assuming the null hypothesis is true. What the *p*-value actually represents is the probability of obtaining the result assuming the null hypothesis is true *if everything in the model is perfectly specified* (which it virtually never is).[157]

Consider the example described above where the employer has 5,100 men and 4,900 women. We saw that the result is significantly different from

---

[155] Griggs v. Duke Power Co., 401 U.S. 424 (1971).

[156] An example of how decreasing the disparity in pass rates *necessarily* increases the disparity in fail rates can be seen in the following example. Assume a test where the average white score is 100, and the average black score is 85, with each group having a standard deviation of 15. (These figures roughly approximate the race difference in performance on IQ tests.) If a minimum score of 100 is required in order to pass, then about half of white test-takers would earn a passing score (since the required score is the same as the group mean). However, only about 16% of blacks would receive a passing score (since the required score is one standard deviation above the black mean). Thus, blacks would have a pass rate only 32% that of whites and a fail rate (84%) that is 1.6 times as great as the white fail rate of 50%. If the employer decides this is too great an impact and lowers the passing score to 85, then 84% of whites would pass, and 50% of blacks would pass, such that the black pass rate would now go up to about 60% of that of whites, but the black fail rate would now be over three times as great as the white fail rate. Suppose that the employer then decides that this is still too great a disparity, so it sets the passing score at 70 (which at two standard deviations below the white mean and one standard deviation below the black mean, indicates "intellectual disability," *see* Hall v. Florida, 572 U.S. __, 134 S. Ct. 1986, 1999 (2014)). Now, about 98% of whites would pass, and 84% of blacks would pass. The black pass rate is now 85 percent of the white pass rate, which would be small enough not to trigger the EEOC's four-fifths rule. But, look what has happened to the disparity in fail rates. The black fail rate is now eight times that of whites. Thus, by increasing the black pass rate as a percentage of the white pass rate from 32% to 85%, the ratio of black to white failures increased from 1.6:1 to 8:1. The only way to avoid a finding of a statistically significant disparity under the *Lew* rationale would be to eliminate large disparities in *both* pass and fail rates, which could be accomplished only by raising the pass rate of both races to 100% or lowering it to zero, either of which courses would completely eliminate any predictive validity that the test might have.

[157] *See* KLINE, *supra* note 14, at 36–38. *See also* Colquhoun, *supra* note 76, at 3 (noting that "[t]he number will be right only if all the assumptions made by the test were true"); Jesper W. Schneider, *Null Hypothesis Significance Tests: A Mix-up of Two Different Theories*, 102 SCIENTOMETRICS 411, 421 (2015) ("When interpreting p values one should not forget that p is a conditional probability, i.e., the probability of the observed data plus more extreme data, *conditional* on H_0 being true, that the sampling method is random, that all distributional requirements are met, that scores are independent and reliable, and that there is no source of error besides sampling error. . . .").

the expected distribution of a 50:50 sex ratio at the 0.05 level. Suppose, though, that men and women are not, in fact, *exactly* equally interested in the jobs. Assume that men are slightly more interested than women, such that the true "expected" ratio is 51:49 rather than 50:50. This is a small enough difference that it would not be perceptible to the naked eye, unlike sex differences in interest in, say, auto mechanics or child care. In other words, based upon naturalistic observation, no one would expect there to be a sex disparity in the jobs, and it is unlikely that sufficient data would exist to show that the true expectation deviated just slightly from 50:50. Yet, our employer with its statistically significant 51:49 split has *exactly* what would be expected by an omniscient statistician but what appears to be a statistically significant shortfall of female employees to everyone else. If the actual split in interest were 52.5:47.5—still a disparity that is not likely to be obvious to the naked eye—and the employer's workforce exactly reflected that difference with 5,250 men and 4,750 women, the disparity would be statistically significant at a level of 0.000001 under an assumption of equal interest. That is, the conclusion would be that there is less than one chance in a million that a randomly selected employer would have such a disparity if the null hypothesis is true.

When one sees that a tiny misspecification of one variable can change what is actually proportional representation into a highly statistically significant shortfall, one can imagine the effects of the crudely specified variables in the typical multiple-regression analysis.[158]

D. The Alternative Hypothesis Is Chosen Casually

Another of the problems with NHST in employment discrimination cases is that the alternative hypothesis is seldom properly specified. We have mostly talked so far about the null hypothesis (H_0), but the null hypothesis is something that one tests in hopes of rejecting it. In NHST, the "alternative hypothesis" (H_1) is what one is really trying to establish. So, the null hypothesis may be that a given drug yields results no better than a placebo, while the alternative hypothesis—which is really why you are doing the testing—is that the drug yields better results. The null hypothesis and the alternative hypothesis should be defined such that they are both mutually exclusive and exhaustive.[159] In those circumstances, rejection of the null hypothesis implies the truth of the alternative hypothesis.

In employment discrimination cases, the null hypothesis is typically that the distribution of an employer's workforce is a product of forces that are

---

[158] *See, e.g.*, Lavin-McEleney v. Marist College, 239 F.3d 476, 478 (2d Cir. 2001) (using only variables of rank, years of service, division, tenure status, and degrees in a compensation-discrimination case involving college faculty).

[159] David Trafimow, *Hypothesis Testing and Theory Evaluation at the Boundaries: Surprising Insights from Bayes's Theorem*, 110 PSYCHOL. REV. 526, 526 (2003).

purely random with respect to race, sex, or some other criterion of interest, such that any differences that do exist are a product of sampling error. The alternative hypothesis is that there are non-random forces operating on the employer's workforce such that members of the relevant groups do not have the same probability of being employed by the employer. This alternative satisfies the requirements of mutual exclusivity and exhaustiveness, but it should really be the beginning, not the end, of the inquiry. Acceptance of this alternative hypothesis does not mean that the employer took race or sex into account in making employment decisions or even that the distribution is directly attributable to employer action at all. As Charles Lambdin has observed:

> [T]he rejection of any nil hypothesis (a null of no difference) supports all research hypotheses predicting an effect, not just yours—and there may be an infinite number of explanations for the effect in question. In fact, all significance here tells you is that you are justified in proceeding to test your research hypothesis, not that your research hypothesis is supported. And yet this fantasy leads to many articles being accepted for publication whenever *p* values are erroneously taken to suggest the research hypothesis is likely correct when the methodology and hypotheses involved are themselves doubtful.[160]

In many cases, the race or sex disparities will be due to race-neutral and sex-neutral reliance on traits that tend to correlate with race or sex. If whites or men tend to have more relevant experience than blacks and women, for example, then, to the extent that prior experience influences the likelihood of hiring, outcomes will not be random with respect to race or sex. In the racial context, racial groups differ substantially in measured IQ.[161] If an employer's practices tend to result in selection of applicants with greater cognitive ability—whether because of reliance on results of standardized tests that are correlated with IQ (as most are) or because many elements of an employee's work history are also correlated with IQ[162]—then the probability

---

[160] Charles Lambdin, *Significance Tests as Sorcery: Science Is Empirical–Significance Tests Are Not*, 22 THEORY & PSYCHOL. 67, 76 (2012).

[161] See *supra* note 149.

[162] Frank L. Schmidt & John Hunter, *General Mental Ability in the World of Work: Occupational Attainment and Job Performance*, 86 J. PERS. & SOC. PSYCHOL. 162 (2004) (finding a strong relationship between general mental ability and job performance (and stronger than between personality and job performance)); Linda S. Gottfredson, *Why g Matters: The Complexity of Everyday Life*, 24 INTELLIGENCE 79 (1997) (reviewing evidence of the importance of intelligence to job performance, especially, but not only, in complex jobs); Nathan R. Kuncel et al., *Academic Performance, Career Potential, Creativity, and Job Performance: Can One Construct Predict Them All?*, 86 J. PERS. & SOC. PSYCHOL. 148 (2004) (finding that tests of general mental ability predict not only academic success but also job performance); John Hunter & Frank L. Schmidt, *Intelligence and Job Performance: Economic and Social Implications*, 2 PSYCHOL., PUB. POL. & LAW, 447 (1996) (intelligence is the strongest determinant of job performance, primarily as a consequence of faster and greater acquisition of job knowledge).

of selection of members of different racial groups is likely to be different.[163]

A distribution that differs from the null hypothesis may not even be the result of any action on the part of the employer at all, but rather because members of the relevant group found the job less attractive than others and were less likely to apply.[164] A job that is either physically or financially risky, for example, is likely to be less appealing to women than it is to men, and a job that has a large social dimension is likely to be more appealing to women than it is to men.[165]

The belief that rejection of the null hypothesis confirms the alternative hypothesis has been labeled "the meaningfulness fallacy," which often reflects two cognitive errors.[166] The first is the belief that a rejection of the null hypothesis in a single study implies that the alternative hypothesis is "proven." The second is the belief that if the *statistical* alternative hypothesis is correct, then the *substantive* alternative hypothesis must also be correct. In the discrimination context, the statistical alternative hypothesis is that chance is not responsible for the disparities in an employer's workforce, while the substantive alternative hypothesis is that the employer has discriminated. As Rex Kline has noted, if the alternative hypothesis is a substantive one, "the work just begins after rejecting H_0," because now the task is to evaluate all of the competing substantive hypotheses that are consistent with the statistical alternative hypothesis, and if alternative explanations cannot be ruled out, "confidence in the original hypothesis must be tempered."[167] The significance test simply does not help in selecting among competing research hypotheses.[168]

Rejection of the null hypothesis in discrimination cases does not distinguish among the myriad scenarios (other competing hypotheses) that might result in a nonrandom workforce. All it does (at most) is eliminate sampling error. Stated that way, the only alternative hypothesis that can be

---

[163] If the disparity is simply due to differences in performance on a standardized test, the proper challenge is based on disparate impact, in which case a racial disparity can be defended by demonstrating that the test is "job related for the position in question and consistent with business necessity." But if the disparity is due to the employer's consideration of a host of more subjective factors that are correlated with IQ—such as, plausibly, education, references from prior employers, apparent job knowledge, or articulateness—then data are unlikely to exist to explain the nonrandom effects of an employer's race-blind selection process.

[164] For a description of sex differences in occupational interest, see *supra* note 150. If good applicant-flow data exist, then those differences can be incorporated into the statistical study, but if no such data exist, as they often do not, then the assumption is likely to be that there is no difference. *See, e.g.*, Catlett v. Missouri Hwy. and Transp. Comm., 828 F.2d 1260, 1266 (8th Cir. 1987), *cert denied*, 485 U.S. 1021 (1988) (comparing the 10 percent of the state's road-maintenance hires that were women with the 48% of women in the general labor force); *cf.* U.S. Bureau of Labor Statistics, *Women in the Labor Force: A Databook*, Table 11, at 43 (2015) (reporting that in 2014, 1.5% of all highway maintenance workers were women).

[165] See *supra* note 150.

[166] KLINE, *supra* note 14, at 100.

[167] *Id.*

[168] Lambdin, *supra* note 160, at 76.

accepted as a consequence of rejection of the null hypothesis is that some action on the part of either employers or potential employees is resulting in a workforce that is not an exact replication of the demographics of the relevant labor force. Discrimination is just one of several possible alternative hypotheses,[169] and there is often no reason to think that it is the most plausible one. Nonetheless, some courts have explicitly placed the burden on employers to show which substantive alternative hypothesis is correct once the statistical alternative hypothesis is concluded to be correct. That is, once sampling error is ruled out, they place the burden on the employer to demonstrate what nondiscriminatory factor is responsible for the disparity.[170] Again, that is contrary to the general rule that the plaintiff bears the burden of persuasion in discrimination cases.

IV. CONCLUSION

Statistical evidence is widely perceived as providing the critical link between a lack of proportional representation in an employer's workforce and an employer's invidious intent. It is true that the greater the disparity between an employer's workforce and the statistically "expected" workforce, the more likely it will usually be that a non-random cause is responsible, though that cause may not be discrimination.[171] What this Article has attempted to establish is that the probability cannot be quantified in the way that it routinely has been in the past, and, in any event, there is no reason to assume that discrimination is the most likely non-random cause of disparities.

Null hypothesis significance testing, relying as it does on *p*-values and labeling of differences as "statistically significant," provides a way, and a relatively convenient one, of assessing liability. However, its relationship to the real world—and especially to what people who use it thinks it does (*i.e.*, provide a probability that the employer's selection was non-random, or even discriminatory)—is attenuated at best. One might easily succumb to the temptation noted by Jacob Cohen to rename NHST "statistical hypothesis inference testing" for its more apt acronym.[172]

Evidence of *p*-values and "statistical significance" should be excluded from discrimination trials. Such evidence may arguably be "relevant" under the extremely low hurdle established by Rule 401 of the

---

[169] *See* Tagatz v. Marquette, 861 F.2d at 1044 ("All that the data show is that there is in all likelihood a *real*, not a spurious, difference between the means of the samples compared. The data do not show that the difference is due to a particular attribute (namely, being Catholic)."); BALDUS & COLE, *supra* note 55, at § 9.02 ("[I]n order to assess the likelihood that chance was the causal factor in a disparate treatment case, it is necessary to assume that all applicants were similarly situated on all relevant qualifications . . . ."); *see generally* Browne, *Beyond Damned Lies*, *supra* note 5, at 503–541.

[170] *See, e.g.*, Adams v. Ameritech Servs., Inc., 231 F.3d 414, 424 (7th Cir. 2000) (noting that once the null hypothesis of random selection was rejected, it was the employer's responsibility to offer alternative explanations).

[171] KAYE & FREEDMAN, *supra* note 17, at 378.

[172] Cohen, *supra* note 8, at 997.

Federal Rules of Evidence—requiring only that the evidence have "any tendency to make a [material] fact more or less probable than it would be without the evidence."[173] However, because of the ease with which *p*-values and statistical significance are misunderstood, and the apparent difficulty—if not impossibility—of overcoming that misunderstanding, they should be barred under Rule 403, which permits the court to exclude evidence "if its probative value is substantially outweighed by a danger of . . . confusing the issues, [or] misleading the jury."[174] Indeed, it seems that misleading of the jury is not just a collateral consequence of the evidence but rather its primary function.

Although it is reassuring to be armed with quantitative studies that purport to establish to several decimal places the probability that the employer was selecting randomly with respect to forbidden criteria, these studies simply do not establish in most cases what they are claimed to establish. If an employer is truly engaging in discrimination as its standard operating procedure, there ought to be disparities that are obvious to the naked eye and numerous flesh-and-blood examples of it.[175] Statistical evidence that is merely consistent with discrimination, but not particularly probative of it, should not serve as the basis for multi-million-dollar judgments. Instead of imposing liability based on a misunderstanding of probabilities, assessment of liability should rest on more complex, subjective human judgments untainted by statistical legerdemain.

[173] FED. R. EVID. 401. Under Rule 402, "[i]rrelevant evidence is not admissible." FED. R. EVID. 402.

[174] FED. R. EVID. 403.

[175] For an extended discussion of ways that the treatment of statistical evidence should be modified, see Browne, *Beyond Damned Lies*, *supra* note 5, at 541–56.