Pepperdine Law Review

Volume 37 | Issue 3 Article 1

3-15-2010

Significant Statistics: The Unwitting policy Making of Mathematically Ignorant Judges

Michael I. Meyerson

William Meyerson

Significant Statistics: The Unwitting Policy Making of Mathematically Ignorant Judges

Michael I. Meyerson & William Meyerson*

I. INTRODUCTION

What could cause a judge to permit prosecutors to highlight the race of criminal defendants when there is no indication that race is relevant to the case?[1] Why would a court require a jury to consider a suspect as likely to be guilty as innocent, even when there is no evidence other than an accusation?[2] And what could possibly induce a judge to permit African-Americans and women to receive lower damages than identically situated white men on the unspoken expectation that racism and sexism will continue for the foreseeable future?[3] These, as well as other, perversions of justice are the direct, though unthinking, result of judicial mathematical illiteracy.

---

* The authors are respectively Professor of Law and Piper & Marbury Faculty Fellow, University of Baltimore School of Law; and Ph.D. candidate in Mathematics at the University of California, Los Angeles; University of Cambridge, Certificate of Advanced Study in Mathematics with Distinction, June 2005; Harvard University, Mathematics, B.A. 2004. We are grateful to Robert H. Lande, Audrey McFarlane, and Max S. Oppenheimer for their insightful comments and suggestions.

1. *See infra* notes 115–69 and accompanying text.
2. *See infra* notes 83–103 and accompanying text.
3. *See infra* notes 170–230 and accompanying text.

The fact that many judges suffer from an "estrangement from, resistance to, and incapacity in mathematics," should not be surprising.[4] This condition, after all, afflicts most lawyers, as it does most Americans.[5] One need not conduct a study to know that "[l]aw students are typically smart people who do not like math."[6] Law professors are of little help to their students, because "legal academics . . . tend not to have a background in, or use, statistical analysis, or . . . are unfamiliar with empirical data collection."[7] Indeed, it is clear that there is a "prevalent (and disgraceful) math-block that afflicts the legal profession."[8]

Mathematical illiteracy is especially worrisome, as the analysis of numbers has become such an important component of the legal system. In particular, "the use of statistical testimony at trial has increased dramatically during the past two decades."[9] Statistical evidence is now an essential element of cases spanning the legal universe. Statistics are regularly used to prove or disprove issues as diverse as causation of injuries in toxic torts cases, breach of contracts, discrimination in employment and voting, DNA identification in criminal and family law, trademark and patent violations, environmental harm, securities fraud, and loss of future earnings.[10]

---

4. Peter A. Coclanis, *History by the Numbers: Why Counting Matters*, 7 MAG. HIST. 5, 8 (1992).

5. *See generally* MARILYN BURNS, MATH: FACING AN AMERICAN PHOBIA (1998).

6. Michael J. Saks, *Legal Policy Analysis and Evaluation*, 44 AM. PSYCHOLOGIST 1110, 1115 (1989).

7. Jeremy A. Blumenthal, *Law and Social Science in the Twenty-first Century*, 12 S. CAL. INTERDISC. L.J. 1, 2 (2002); *see also* Richard A. Posner, *An Economic Approach to the Law of Evidence*, 51 STAN. L. REV. 1477, 1479 (1999) (stating "legal education itself (alas) 'produces no improvement in the ability to apply the statistical and methodological rules of the probabilistic sciences to either scientific studies or everyday-life events.'" (quoting Darrin R. Lehman et al., *The Effects of Graduate Training on Reasoning: Formal Discipline and Thinking About Everyday-Life Events*, 43 AM. PSYCHOLOGIST 431, 440 (1988))).

8. Jonathan J. Koehler, *The Probity/Policy Distinction in the Statistical Evidence Debate*, 66 TUL. L. REV. 141, 148–49 (1991) (quoting Richard A. Posner, *The Decline of Law as an Autonomous Discipline: 1962-1987*, 100 HARV. L. REV. 761, 778 (1987)). As Justice Breyer has noted, "judges are not scientists and do not have the scientific training that can facilitate the making of such decisions." Gen. Elec. Co. v. Joiner, 522 U.S. 136, 148 (1997) (Breyer, J., concurring).

9. Koehler, *supra* note 8, at 141.; *see also Gen. Elec. Co.*, 522 U.S. at 149 (Breyer, J., concurring) (stating "cases presenting significant science-related issues have increased in number"); FEDERAL COURTS STUDY COMMITTEE, JUDICIAL CONFERENCE OF THE UNITED STATES, REPORT OF THE FEDERAL COURTS STUDY COMMITTEE 97 (1990) ("Economic, statistical, technological, and natural and social scientific data are becoming increasingly important in both routine and complex litigation.").

10. David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, *in* FEDERAL JUSTICE CENTER, REFERENCE MANUAL ON SCIENTIFIC EVIDENCE, 83, 85 (2d ed. 2000) [hereinafter Kaye & Freedman, *Reference Guide on Statistics*] ("Statistical assessments are prominent in many kinds of cases, ranging from antitrust to voting rights. Statistical reasoning can be crucial to the interpretation of psychological tests, toxicological and epidemiological studies, disparate treatment of employees, and DNA fingerprinting; this list could easily be extended."); *see also, e.g.*, Anderson v. Westinghouse Savannah River Co., 406 F.3d 248 (4th Cir. 2005) (racial discrimination in

The Supreme Court has declared that in all such cases, federal courts have a "general 'gatekeeping' obligation."[11] Beginning with its 1993 decision in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[12] the Court has declared that the Federal Rules of Evidence "assign to the trial judge the task of ensuring that an expert's testimony both rests on a reliable foundation and is relevant to the task at hand."[13] Put bluntly, the Court has told judges that they must ascertain the difference between "good science"[14] and "junk science."[15]

This evaluation appears to be a substantially more subtle task than that formerly performed by federal judges. Under the previous regimen, pursuant to the so-called *Frye* test, courts would admit statistical and other technical evidence if it could be shown that such evidence was derived from a well-recognized scientific principle or discovery which had "gained general acceptance in the particular field in which it belongs."[16]

---

employment); Marvin Lumber & Cedar Co. v. PPG Indus., Inc., 401 F.3d 901 (8th Cir. 2005) (breach of warranty claim); Currier v. United Techs. Corp., 393 F.3d 246 (1st Cir. 2004) (age discrimination claim); Citizens Fin. Group, Inc. v. Citizens Nat'l Bank of Evans City, 383 F.3d 110 (3d Cir. 2004) (trademark "reverse confusion" case); United States v. Blaine County, Montana, 363 F.3d 897 (9th Cir. 2004) (claiming that at-large election system diluted vote of Native Americans in violation of the Voting Rights Act); *In re* Hanford Nuclear Reservation Litig., 292 F.3d 1124 (9th Cir. 2002) (causation in toxic tort personal injury case); Boncher *ex rel* Boncher v. Brown County, 272 F.3d 484 (7th Cir. 2001) (Due Process violated by lax prison security); Chavez v. Ill. State Police, 251 F.3d 612 (7th Cir. 2001) (ethnic profiling in traffic stops); United States v. Wright, 215 F.3d 1020 (9th Cir. 2000) (DNA match of bank robber); Lehocky v. Tidel Techs., Inc., 220 F.R.D. 491 (S.D. Tex. 2004) (securities fraud); Everett v. Everett, 201 Cal. Rptr. 351 (1984) (paternity suit).

11. Kumho Tire Co. v. Carmichael, 526 U.S. 137, 141 (1999); *see also* Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579, 589 n.7 (1993) (referring to the judge having "gatekeeping responsibility"). Judges are to perform this gatekeeping function for testimony based on "scientific" "technical" and "other specialized" knowledge. *Kumho Tire Co.*, 526 U.S. at 141.

12. 509 U.S. 579 (1993).

13. *Id.* at 597. Rule 702 of the Federal Rules of Evidence states: "If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise." FED. R. EVID. 702.

14. *Daubert*, 509 U.S. at 593.

15. Gen. Elec. Co. v. Joiner, 522 U.S. 136, 153 (1997) (Stevens, J., dissenting).

16. Frye v. United States, 293 F. 1013, 1014 (1923). While twenty-five states (Alaska, Arkansas, Colorado, Connecticut, Delaware, Idaho, Indiana, Iowa, Kentucky, Louisiana, Maine, Montana, New Mexico, North Carolina, Ohio, Oklahoma, Oregon, Rhode Island, South Carolina, South Dakota, Tennessee, Texas, Vermont, West Virginia, and Wyoming) have expressly adopted the *Daubert* standard or a similar test, at least fifteen states (Arizona, California, Florida, Illinois, Kansas, Maryland, Michigan, Minnesota, Mississippi, Missouri, Nebraska, New York, North Dakota, Pennsylvania, and Washington) and the District of Columbia continue to utilize the *Frye* standard. *See* Alice B. Lustre, Annotation, *Post-Daubert Standards for Admissibility of Scientific and Other Expert Evidence in State Courts*, 90 A.L.R.5th 453, 454–55 (2008).

The Supreme Court has explained that the *Daubert* test is less "rigid" than *Frye*.[17] The Court stated that, without the exclusive emphasis on "general acceptance," federal judges would be able "to admit a somewhat broader range of scientific testimony than would have been admissible under *Frye* . . . ."[18] Nonetheless, the Court has stressed that trial judges must still screen evidence to determine its "scientific validity."[19] According to the Court, this inquiry is a "flexible" one.[20] The Court also proposed several factors that might "bear on" this examination.[21] These factors include:

> —Whether a "theory or technique . . . can be (and has been) tested";
> —Whether it "has been subjected to peer review and publication";
> —Whether, in respect to a particular technique, there is a high "known or potential rate of error" and whether there are "standards controlling the technique's operation;" and
> —Whether the theory or technique enjoys "general acceptance" within a "relevant scientific community."[22]

Daubert's guidance has not provided much comfort for some judges who have complained that they have been placed in the uncomfortable position of evaluating the quality of scientific and statistical evidence far beyond their own fields of expertise:

> [T]hough we are largely untrained in science and certainly no match for any of the witnesses whose testimony we are reviewing, it is our responsibility to determine whether those experts' proposed testimony amounts to "scientific knowledge," constitutes "good science," and was "derived by the scientific method."
>
> . . . .
>
> Our responsibility, then, unless we badly misread the Supreme Court's opinion, is to resolve disputes among respected, well-credentialed scientists about matters squarely within their expertise, in areas where there is no scientific consensus as to what is and what is not "good science," and occasionally to reject such expert testimony because it was not "derived by the scientific method."

---

17. *Daubert*, 509 U.S. at 588.
18. *Gen. Elec. Co.*, 522 U.S. at 142.
19. *Daubert*, 509 U.S. at 594.
20. Kumho Tire Co. v. Carmichael, 526 U.S. 137, 141, 150 (1999).
21. *Id.* at 149.
22. *Id.* at 149–50 (summarizing *Daubert*, 509 U.S. at 592–94) (internal quotation marks omitted).

> Mindful of our position in the hierarchy of the federal judiciary, we take a deep breath and proceed with this heady task.[23]

Such admirable humility might not be necessary were judges truly to limit themselves to keeping out blatantly "junk science," such as "the testimony of a phrenologist who would purport to prove a defendant's future dangerousness based on the contours of the defendant's skull."[24] Similarly easy evidence to exclude would be, as the Court explained, "theories grounded in any so-called generally accepted principles of astrology or necromancy."[25]

Even obviously bad mathematics should be readily observable to the average judge. Consider the case of *Boncher ex rel Boncher v. Brown County*.[26] In the course of a lawsuit, alleging that a jail had been deliberately indifferent to the risk of prisoner suicide, a criminologist testified that the risk was "particularly acute" because there had been five suicides at the jail during the previous five years.[27] But how "big" a risk does the number "five" convey? As the court noted, the simple number "five" does not disclose whether the risk was "acute" unless we also know the size of the prison population, the rate of suicides in other prisons, and the rate of suicides in the general geographic area from which the jail draws its inmates.[28] After all, five suicides in a small prison population obviously demonstrate a far greater level of risk than were that same number to occur in a very large population. The criminologist's numeric evidence was indeed "useless" and properly excluded under *Daubert*.[29]

On more subtle points of mathematics or science, it would be appropriate for judges to turn to experts to help identify whether proposed evidence is "good science." As the editors of the New England Journal of Medicine suggested, "Judges should be strongly encouraged to make greater use of their inherent authority . . . to appoint experts."[30]

---

23. Daubert v. Merrell Dow Pharms., Inc., 43 F.3d 1311, 1316 (9th Cir. 1995).

24. *Gen. Elec. Co. v. Joiner*, 522 U.S. 136, 153 n.6 (1997) (Stevens, J., dissenting).

25. *Kumho Tire Co.*, 526 U.S. at 151. "Necromancy" is the "practice of supposedly communicating with the spirits of the dead in order to predict the future." THE AMERICAN HERITAGE DICTIONARY OF THE ENGLISH LANGUAGE (4th ed. 2004).

26. 272 F.3d 484 (7th Cir. 2001).

27. *Id.* at 486.

28. *Id.* at 486–87.

29. *Id.* at 486.

30. Gen. Elec. Co. v. Joiner, 522 U.S. 136, 149–50 (1997) (Breyer, J., concurring) (citation omitted).

But non-legal experts have been permitted to expand their role to the point where they are now making normative policy decisions. Sometimes this has happened when courts have confused scientific validity with logical relevance; the science and mathematics may be undisputed, but the question really involves the logical and legal connection of the numbers to the legal point being proven.[31] At other times, courts have been unable to distinguish the methods of the scientist and statistician from the values of their disciplines. While those who use these methods must respect the methodology of the disciplines from which they arise, we must also recognize that the values of those disciplines often differ markedly from those of the legal system.[32] As the Carnegie Commission report, *Science and Technology in Judicial Decision Making*, observed: "In the courts, scientific knowledge must inform the choice, but abdication to the scientist is incompatible with the judge's responsibility to decide the law."[33] In other words, while striving to avoid accepting "junk science" into evidence, too many judges have permitted statisticians and others to allow "junk law" into the courts.

Many have assumed the problem is that judges "lacked the scientific literacy" to evaluate evidence properly.[34] Thus, there are complaints that "[t]hose of a 'scientific' bent certainly can take issue with whether the judges and lawyers have the education or training to engage in 'scientific' testing . . . ."[35] The critical judicial deficit, however, is not in science, but in mathematics. Too many judges do not "speak math" and do not understand what numbers communicate. They also fail to see that the meaning to be given to mathematical results is frequently not a matter of scientific necessity, but a reflection of specific value judgments. By ignoring those judgments that are inherent in the mathematical choices, judges have acquiesced to values that are at odds with our system of justice.

This Article will explore several areas in which judges, hampered by their mathematical ignorance, have permitted numerical analysis to subvert the goals of our legal system. In Part II, I will examine the perversion of the presumption of innocence in paternity cases, where courts make the counter-

---

31. David H. Kaye, *Rounding Up the Usual Suspects: A Legal and Logical Analysis of DNA Trawling Cases*, 87 N.C. L. REV. 425, 431 (2009) [hereinafter Kaye, *Rounding Up the Usual Suspects*]; *see also infra* notes 113–14 and accompanying text.

32. *See infra* notes 231–42 and accompanying text.

33. CARNEGIE COMMISSION ON SCIENCE, TECHNOLOGY, AND GOVERNMENT, SCIENCE AND TECHNOLOGY IN JUDICIAL DECISION MAKING: CREATING OPPORTUNITIES AND MEETING CHALLENGES 27 (1993).

34. Sophia I. Gatowski et al., *Asking the Gatekeepers: A National Survey of Judges on Judging Expert Evidence in a Post-*Daubert *World*, 25 LAW & HUM. BEHAV. 433, 433 (2001); *see also* Paul S. Miller et al., Daubert *and the Need for Judicial Scientific Literacy*, 77 JUDICATURE 254, 254 (1994).

35. United States v. Cline, 188 F. Supp. 2d 1287, 1294 (D. Kan. 2002).

factual assumption that regardless of the evidence, prior to DNA testing a suspect has a fifty-fifty chance of being the father.[36] In Part III, I will explore the unnecessary injection of race into trials involving the statistics of DNA matching, even when race is entirely irrelevant to the particular case.[37] Next, in Part IV, I will discuss how courts use race- and gender-based statistics to reduce damages in tort cases for women and racial minorities, and silently assert that past racism and sexism will continue.[38] In the final section, I will examine how judges have improperly allocated the risk of error in cases such as securities fraud, so as to reward those who have attempted to manipulate stock prices illegally.[39]

II. SNAKE EYES AND THE POWER OF NUMBERS

To understand both the uses and abuses of statistical evidence, I will present a simple story, a murder mystery, called *Snake Eyes*.[40] I will then demonstrate the limits of pure statistical analysis and the way that courts have permitted statisticians and scientists to warp the legal process and pervert traditional legal values.

Snake Eyes

> Victor is an elderly millionaire who has decided to bring his family together to give away his possessions, which range in value from one extremely rare antique to several mundane items. He invites his eleven closest relatives to his house. He seats them at a long table and tells them that they will be rolling dice to determine the order in which they will select their respective gifts, with the highest roll choosing first. Victor opens a fresh set of dice from the Trustworthy Dice Company and hands them to his guests. Each guest simultaneously places two dice in a dice cup, tosses the dice, looks at the result, and covers their dice with the cup.
>
> Sitting at one end of the table are Al and Dennis. Al, immediately after looking at his dice, runs out of the house. All the other guests race to the window to watch Al get into his car and

36. *See infra* notes 40–114 and accompanying text.
37. *See infra* notes 115–69 and accompanying text.
38. *See infra* notes 170–230 and accompanying text.
39. *See infra* notes 231–425 and accompanying text.
40. The term "snake eyes" refers to a roll of two dice which results in each die landing with one spot face up. MICROSOFT ENCARTA COLLEGE DICTIONARY 1366 (2001).

drive away. Suddenly, they hear a loud sound and turn around to see Victor on the floor, bludgeoned to death with a candlestick. By Victor's body is a note saying, "I killed him because I rolled snake eyes."

The guests rush back to their seats and find that Victor's fall had knocked all of the dice cups onto the floor, scattering all of the dice except for Al's and Dennis's. Both dice cups are lifted, revealing that each had rolled two "ones." Dennis concedes that indeed he had rolled "snake eyes," but he denies having killed Victor.

Assuming the note was accurate, the dice were fair, and that there is no other evidence, what can statistics tell us about the identity of the murderer? Can it tell us how likely it is that Dennis is the murderer?

We can calculate easily the probability of Dennis rolling the murderer's two "ones" if he were innocent. The probability that Dennis matched the murderer's roll by pure random chance is one in thirty-six, which is about 2.8%.[41] What that tells us is that Dennis's roll matched that of the murderer, and the probability of a random match is 2.8%. Another way to think about this is that if Dennis were not the murderer, the probability of seeing this match is 2.8%.[42]

But that does not tell us what we want to know, which is the probability that Dennis is the murderer. It is incorrect to say that: (1) Because the probability of Dennis matching the murderer's snake eyes as a matter of random chance was only 2.8%, then (2) The probability of Dennis not being the murderer, given that match, is the same small 2.8%.

This error in sentence (2) is called the "prosecutor's fallacy" because it incorrectly reverses events in a conditional probability to create a direct statement about the defendant's probability of guilt that is not implied by the evidence.[43] In logical reasoning, such an error is called "transposing the conditional."[44] It is the same mistake as saying: "Because lawyers tend to be literate people, literate people tend to be lawyers."

To understand what other information is needed to calculate the probability of guilt, we have to keep in mind that the likelihood of Dennis's guilt depends in large measure on whether *other* people could have

41. The probability of obtaining a "six" on one die is one in six. Because each die's outcome is independent of the other, the probability of obtaining a "six" on two dice is calculated by multiplying the probability for obtaining a "six" on each: 1/6 x 1/6 = 1/36. HENRY E. KLUGH, STATISTICS: THE ESSENTIALS FOR RESEARCH 152 (3d. ed. 1986).

42. Roger C. Park & Michael J. Saks, *Evidence Scholarship Reconsidered: Results of the Interdisciplinary Turn*, 47 B.C. L. REV. 949, 989 (2006).

43. *See* WOJTEK J. KRZANOWSKI, STATISTICAL PRINCIPLES AND TECHNIQUES IN SCIENTIFIC AND SOCIAL INVESTIGATIONS 18 (2007); *see also* William C. Thompson, *DNA Evidence in the O.J. Simpson Trial*, 67 U. COLO. L. REV. 827, 850 (1996).

44. Park & Saks, *supra* note 42, at 988.

committed the crime. If no one else in the room rolled snake eyes, then there is no other suspect, and we can be 100% certain that Dennis is guilty. If one other person in the room also rolled snake eyes, however, the probability that Dennis is the murderer is fifty percent, because it would be equally likely for either to be guilty.[45] The key question, then, is not how unlikely it was for Dennis to have rolled the telltale snake eyes, but how many other people in the room also did.

Mathematically, that means we need to account for both the probability that other people at the gathering rolled snake eyes and the probability that Dennis was indeed the murderer despite the existence of others who matched the evidence. In this story, there were ten people at the party who could have committed the crime because Al was not in the house when the murder happened. In such a case, the probability of Dennis's guilt is 88.4%.[46] But, if there were fewer suspects, say only five, the likelihood of finding other suspects rolling snake eyes goes down, and the probability of Dennis's guilt would increase to 94.6%. On the other hand, if there were many more members of the family, say 100 other relatives in the room, there would then be many more possible suspects, and the probability of Dennis's guilt would drop to 33.8%. The probability of Dennis having rolled snake eyes has not changed with each scenario, but the probability that he is the murderer—the only issue we care about—has varied greatly depending on how many other possible suspects there are.

It was not difficult to calculate the extent of this variation for these examples because the number of other possible suspects was fixed and known. One problem with applying this approach in the real world is that we usually do not know how many other possible suspects there are. If a murder occurs on a street in West Baltimore, the number of possible suspects would turn, in part, on whether the universe of suspects includes only those who live in the neighborhood, residents of the city, or all who might conceivably have visited Baltimore that day.[47] Each choice will lead

45. Mathematically, if n people rolled snake eyes, the probability of a given snake eyes roller being the murderer is $1/n$.

46. The probability that a person rolling snake eyes is guilty can be expressed mathematically as:

$$\sum_{J=1}^{N} [P((J-1) \text{ other snake eyes rolls}) \times (1/J)]$$

with "N" equaling the total number of potential suspects and "J" equalling the differing number who could have rolled snake eyes.

47. *See generally The Wire* (HBO television series 2002–2008).

to a different probability that a particular suspect is guilty. Accordingly, how we answer that unanswerable question will determine the result of any statistical analysis. This dilemma has been called "the problem of reference classes."[48]

Resolving the ultimate question of the probability of a particular suspect's guilt requires us to choose the particular population group with which to compare our evidence. Unfortunately, there are "an infinite number of reference classes, the boundary conditions of which can be gerrymandered in countless ways . . . . [and] nothing in the natural world privileges or picks out one of the classes as the right one . . . ."[49] To do the math, we must choose the appropriate reference class, but we are then making a subjective judgment not mandated by objective analysis.[50]

We face, therefore, one of the sad realities of statistics. The easiest number to calculate—the probability that the defendant matches particular evidence—does not give us the information we really want. Even if it were very unlikely that the defendant matched particularly damning evidence, the numbers would not tell us how likely it was that the evidence came from the defendant.

We encounter this same problem when trying to determine paternity using DNA.[51] The facts of the New Jersey case, *State v. Spann*,[52] illustrate the issue. In *Spann*, the defendant, a corrections officer at the county jail, was charged with sexually assaulting an inmate.[53] To prove that the defendant had had sex with the victim, the prosecution wanted to prove that the defendant was the father of the victim's child.[54]

Blood tests were entered into evidence.[55] They showed that the child had phenotypes A2, A28, B45, and B53.[56] This corresponds to two genes,

---

48. Ronald J. Allen & Michael S. Pardo, *The Problematic Value of Mathematical Models of Evidence*, 36 J. LEGAL STUD. 107, 109 (2007).

49. *Id.* at 112.

50. *See, e.g.*, *id.* ("[O]ur interests in the various inferences they generate pick out certain classes as more or less relevant.").

51. DNA (deoxyribonucleic acid) is a molecule that contains the genetic information for all living things. David H. Kaye & George F. Sensabaugh, Jr., *Reference Guide on DNA Evidence*, *in* FEDERAL JUSTICE CENTER, REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 485, 487 (2d ed. 2000) [hereinafter Kaye & Sensabaugh, *Reference Guide on DNA Evidence*]. Most human DNA can be found in our chromosomes. *Id.* at 491. A fertilized human egg has twenty-three pairs of chromosomes, with each parent contributing half. NATIONAL RESEARCH COUNCIL, THE EVALUATION OF FORENSIC DNA EVIDENCE 60 (1996) [hereinafter *NRC II*]. Each chromosome contains many thousands of genes, which are segments of DNA which have specific functions, determining all the physical traits that we inherit from our biological parents. Kaye & Sensabaugh, *Reference Guide on DNA Evidence*, *supra* at 492. Important traits are usually the product of the relationship between many different genes. *Id.*

52. 617 A.2d 247 (N.J. 1993).

53. *Id.* at 248.

54. *Id.* at 249.

55. *Id.*

located next to one another on the same chromosome.[57] Of these four phenotypes, a child receives two sets of A and B genes, one from the mother and one from the father. The mother had HLA types A28, A30, B53, and B61, which meant that the child's "A2, B45" set had to come from the father. Thus, if a person did not have the A2 and B45 genes, that person could not be the father. (Analogously, in the story *Snake Eyes*, if one guest, Iris, had rolled two "fives", we would know that she was definitely not the murderer.) Accordingly, a blood test can eliminate a suspect as a possible father. As the Supreme Court has noted, "It is a negative rather than an affirmative test with the potential to scientifically exclude the paternity of a falsely accused putative father."[58]

Just as in *Snake Eyes*, where one could calculate the probability of Dennis matching the murderer's snake eyes as a matter of random chance, it is possible in *Spann* to calculate the probability of the defendant matching the child's genes as a matter of random chance. Genetic tables showed that only one percent of the relevant male population had the requisite blood and tissue type.[59] But, as in *Snake Eyes* above, the fact that there is a small probability that the defendant would match the child's genes as a matter of random chance does not establish a similarly small probability that the defendant was the actual source of the child's genes.

What that one percent figure does tell us is what is known as the "probability of exclusion."[60] Ninety-nine percent of the relevant male population can be excluded from suspicion.[61] That still leaves one percent. In a male population of 100,000, that would mean 1,000 people had not been excluded. The probability of exclusion also cannot not tell us who among those 1,000 is most likely to be the father.

---

56. *Id.* at 250 n.1. A "phenotype" has been defined as "[a] trait, such as eye color or blood group, resulting from a genotype." Kaye & Sensabaugh, *Reference Guide on DNA Evidence*, *supra* note 51, at 572.

57. *Id.*

58. Little v. Streater, 452 U.S. 1, 7 (1981) (holding that to deny blood grouping tests in a paternity suit because of defendant's lack of financial resources violated due process). In *Spann*, the defendant's phenotype was A2, A28, B35, and B45. *Spann*, 617 A.2d at 250 n.1. Because he had the A2 and B45 genes, he could not be excluded as the possible source of the child's genes.

59. *Spann*, 617 A.2d at 250 n.1.

60. Robert W. Peterson, *A Few Things You Should Know About Paternity Tests (But Were Afraid To Ask)*, 22 SANTA CLARA L. REV. 667, 680 (1982).

61. *Spann*, 617 A.2d at 250 n.1. Analogously, in *Snake Eyes*, the probability of exclusion (the probability of rolling anything other than two "1s") was thirty-five in thirty-six, or approximately 97.2%.

A statistic that is related to the probability of exclusion, which can also be derived from blood tests, is called the Paternity Index (PI).[62] The PI compares the probability that the genetic makeup of the child could result from the mating of the mother and a particular suspect with the probability that it could result from the mating of the mother with some person randomly selected from the general population.[63] The less frequently a particular genetic marker appears in the general population, the lower the probability that a person chosen at random could be the father, and hence the greater the PI.[64]

Note that the PI, like the probability of exclusion, still does not tell us how likely it is that a particular suspect is the father. For example, consider the following example drawn from a classic article by Professor David Kaye.[65]

Assume that the probability a defendant in a paternity suit would transmit the particular genes in question is 0.12, and that the probability for a randomly selected man was 0.0062. The PI for the defendant would be obtained by dividing his probability by that of the randomly selected man, which would equal 19.4. A PI of 19.4 means that someone with the defendant's exact genetic makeup would produce (with the mother) a child with the requisite phenotype more than nineteen times as frequently as would a randomly selected man.

Yet, we still do not know the probability that the defendant is the father. If there were a relevant population of 100,000 men, we might expect that approximately 620 men other than the defendant would also be capable of transmitting the genes that created the child in question. Based on the PI, that would mean that the probability that the defendant was the father was only a miniscule 0.019%.[66]

In the real world, however, we cannot assume that, just because 620 people have the same matching genetic profile as the defendant, "everyone

---

62. ANDREI SEMIKHODSKII, DEALING WITH DNA EVIDENCE: A LEGAL GUIDE 75–77 (2007). Sometimes, PI refers to a single DNA marker. *Id.* at 76–77. By multiplying the PI for several different markers, we can calculate a Combined Paternity Index (CPI), which is the value usually entered into evidence. *Id.* Because courts typically, simply refer to the PI rather than CPI, *see, e.g.*, Griffith v. State, 976 S.W.2d 241, 243 (Tex. App. 1998), I will use PI instead of CPI in this Article.

63. SEMIKHODSKII, *supra* note 62, at 76.

64. The PI tends to be based on more genetic information than the probability of exclusion because it incorporates not just the existence of particular genetic markers, but also the fact that men with some sets of markers are more likely to transmit the particular genes than men with other sets. D.H. Kaye, *The Probability of an Ultimate Issue: The Strange Cases of Paternity Testing*, 75 IOWA L. REV. 75, 91 (1989).

65. *See id.* at 89–94.

66. If all 100,000 men were seen as having the same opportunity to be the father as the defendant, using Bayesean analysis, *see infra* notes 69–80, the prior odds would be 1 to 99,999, and the posterior odds for the defendant's paternity would be 19.4 to 99,980.6, which equals approximately 0.019%. *See* Kaye, *supra* note 64, at 94.

[is] equally likely to be guilty."[67] This has been termed the "defendant's fallacy"; it uses the numbers to make guilt appear unlikely while ignoring all of the other evidence—such as how many people actually knew the mother, how many of those ever had the opportunity to have sexual relations with her at a time when she was able to conceive, and how many men were infertile—all of which would exclude some people and render others more plausible.[68]

There is a mathematical solution to this problem, but it is one in which the legal system's mathematical ignorance has led to a very disturbing trend. The solution begins with a mathematical formula known as Bayes' Theorem.[69] On its most basic level, the formula is nothing more than a mathematical way of representing how we incorporate new information into our reasoning: "When an observer receives new evidence relevant to the truth of the proposition at issue, she adjusts her probability assessment to take that evidence into account."[70]

For example, suppose a new restaurant opens and I read an excellent review. I would think it is likely that this is a good restaurant. Then, assume that a friend, whose tastes I trust, tells me she ate at the restaurant and that the food was terrible. Obviously my assessment of the probability that the restaurant is good will change with this new information. The degree to which my assessment changes will depend on how much I value the opinion of both the restaurant reviewer and my friend.

That reality is captured by Bayes' Theorem. To utilize this theorem, we need to know that the Bayesian analysis of evidence relies on six concepts: hypothesis, information, prior probability, likelihood, likelihood ratio, and posterior probability.[71]

A *hypothesis* could be thought of as a theory of the case; it is an answer to the question "what happened?" There will always be a "main" hypothesis, the theory being considered. In paternity litigation, the main hypothesis is that the suspect is the father.

---

67. SEMIKHODSKII, *supra* note 62, at 116.

68. *Id.*

69. Richard D. Friedman, *A Presumption of Innocence, Not of Even Odds*, 52 STAN. L. REV. 873, 875 (2000) ("Bayes' Theorem posits that the posterior odds of the proposition equal the prior odds times the likelihood ratio.").

70. *Id.* at 874–75.

71. *See, e.g.*, Kaye, *Rounding Up the Usual Suspects*, *supra* note 31, at 463; Friedman, *supra* note 69, at 875.

Information refers to everything we know. Our aim is to determine the probability that our main hypothesis is true *after* taking into account new information.[72] In the paternity case, the new information is the DNA match.

Before the new information is obtained, a hypothesis has a *prior probability* of being true. Prior probability refers to the probability that a particular hypothesis is true based on everything we knew before the arrival of the new information.[73]

Likelihood refers to the probability of having obtained the new information under the assumption that a particular hypothesis is true.[74] For example, assuming the suspect is the father (our main hypothesis), what is the probability he will match the DNA (our new evidence)? Because we know that the father definitely matched the DNA, that probability is one.[75] Thus, we would say the likelihood is one.

To calculate the *likelihood ratio*, one must divide the likelihood by the probability that the event would occur by random chance.[76] The likelihood ratio shows the effect the new evidence has on our hypothesis. As Richard Friedman notes:

> A likelihood ratio greater than 1 means that the proposition appears more probable in light of the new evidence; a likelihood ratio less than 1 means that the new evidence makes the proposition appear less probable; and a likelihood ratio of precisely 1 means that the new evidence leaves the probability unchanged.[77]

Finally we have our goal, the *posterior probability*, which is the probability of our main hypothesis being true after we have obtained the new information.[78] (Note that this is the transposition of the likelihood.) The posterior probability in a paternity suit is the probability that the suspect is the father after we have the DNA match.

There are many different ways to express Bayes' Theorem.[79] For our purposes, we can use the following:

---

72. Friedman, *supra* note 69, at 874–75.
73. Kaye, *Rounding Up the Usual Suspects*, *supra* note 31, at 463.
74. *See id.*at 464.
75. *Id.*
76. *See* Friedman, *supra* note 69, at 875.
77. *Id.*
78. *Id.*
79. Another way in which Bayes' Theorem is presented is:

$$P(A|B) = P(A) \times \frac{P(B|A)}{(P(B))}$$

$P(A)$ means the probability that event A would occur; $P(B)$ means the probability that event B would occur; $P(A|B)$ means the probability of A given that B has occurred; and $P(B|A)$ means the

Posterior probability = Prior probability x Likelihood ratio of hypothesis

Because the PI tells us the same information as the likelihood ratio, we can write Bayes' formula for paternity litigation as:

Probability suspect = Prior probability x Paternity Index that the suspect is the father

In other words, the probability that the defendant is the father, generally known as the "probability of paternity,"[80] equals the PI multiplied by the prior probability of the defendant being the father. Because we have already seen that the PI can be calculated readily, the only other issue is how to calculate prior probability.

In simplistic cases, like the one described above in *Snake Eyes*, this prior probability is not difficult to calculate. If there were ten guests, prior to knowing the results of rolling the dice, each had an equal chance of being the murderer so that the prior probability that Dennis, or any of the others, was the murderer was one in ten, which equals ten percent. If we already knew that Iris had rolled two "5s", she would be eliminated as a possibility, and the prior probability that Dennis was the murderer would increase to one in nine, or approximately 11.1%.

The real world, of course, is far messier. In the paternity context, for example, the prior probability of the suspect being the father depends both on whether he had sexual relations with the mother and on how many other men she had sexual relations with, as well as the timing of each sexual encounter. Those facts will rarely be known by the fact finder with any degree of certainty, but instead will be conveyed by a wide range of incomplete, uncertain, and often disputed pieces of evidence. Accordingly, any determination of "prior probability"[81] will necessarily be based on the subjective judgment of the fact finder. The prior probability that the suspect is the father will be, in other words, an imprecise approximation, rather than a nice, neat number. The problem that then arises is that Bayes' Theorem is

probability of *B* given that *A* has occurred. $P(A|B)$ is therefore the transposition of $P(B|A)$; *see also, e.g.*, JEFF GILL, BAYESIAN METHODS: A SOCIAL AND BEHAVIORAL SCIENCES APPROACH 7 (2002) ("Bayes' law . . . . is really a device for 'inverting' conditional probabilities.").

80. Friedman, *supra* note 69, at 881 n.22.

81. *See* Kaye, *supra* note 31, at 465.

no longer usable; one cannot do the calculation with only a subjective sense of what might have happened.[82]

Without Bayes' formula, all that is left is the PI, which can tell us how much of the population to exclude, but fails to answer the critical question of how likely it is that the suspect is the father. There is nothing inherently wrong with such a situation. In many trials, jurors are given information that excludes large portions of the population, and must then figure out if the defendant, who was not excluded, is the guilty party. For example, a jury may be shown photographic evidence that the robber was a six-foot-six inch Caucasian male. That would exclude all who do not meet that description, but even if the defendant meets that description, the prosecutor would still need to show more evidence (such as that the stolen goods were found in the defendant's possession) to obtain a conviction.

The power of numbers and the promise of an objective determination, however, have blinded many judges and legislators when it comes to paternity testing. In courts throughout the country, the probability of paternity is calculated using Bayes' Theorem by taking the PI and inserting "a standard prior probability of .5 regardless of any other factors, which indicates a fifty percent chance that the alleged father actually had sexual intercourse with the mother."[83]

What the use of a prior probability of .5 means is that genetic experts present to a jury the "probability of paternity" as a fixed number on the unproven (and often unspoken) assumption that, prior to the genetic testing, there was a fifty-fifty chance that the suspect was the father.[84] Courts have justified this use of a fictitious number on the dubious grounds that, "[t]he 50/50 assumption was completely neutral."[85] This claim that this is a neutral probability is based on the argument that "a prior probability of .5 assumes that the defendant is just as *not* likely the father of the child as it assumes he is the father."[86]

While fifty-fifty may appear fair at first glance, the fifty-fifty assumption of prior probability is demonstrably not a neutral assumption. To see why, let us return to *Snake Eyes*.[87] If Dennis had been accused of

---

82. For a discussion of the impossibility of turning subjective belief into a concrete number, see *infra* text accompanying notes 102–05.

83. Butcher v. Kentucky, 96 S.W.3d 3, 7 (Ky. 2002).

84. *See id.*

85. Davis v. State, 476 N.E.2d 127, 138 (Ind. Ct. App. 1985).

86. Griffith v. State, 976 S.W.2d 241, 250 (Tex. App. 1998); *see also* Brown v. Smith, 526 S.E.2d 686, 689 (N.C. Ct. App. 2000) ("A neutral assessment of the non-genetic evidence would result in a prior probability of 0.5. This would give equal weight to paternity and non-paternity from a non-genetic aspect."); *Butcher*, 96 S.W.3d at 9 ("[A] .5 prior probability is neutral, neither assuming nor denying that intercourse has taken place between the mother of the child and the alleged father.").

87. *See supra* text accompanying note 40.

being the murderer prior to the disclosure of his dice roll, it would have been a tremendous injustice to assume that he was equally as likely to be the murderer as not. There were, after all, nine other guests with an equal chance of being guilty. Thus, a fifty-fifty assumption would be tantamount to saying that Dennis was as likely to be the murderer as everyone else combined. This is hardly a fair or accurate statement, as the odds of guilt were not fifty-fifty, but actually nine-to-one.

One can also recognize this concept by looking at sports betting in competitions with numerous contestants. For example, when there are sixty-five teams in the NCAA Men's Basketball tournament, it would be ludicrous to say that it is just as likely for one team to win as all the others, that the odds are fifty-fifty. Indeed, for the 2008 tournament, even the favorite, UCLA, was given odds by bookmakers of seven-to-two, while long shot Coppin State was given odds of 2500-to-1.[88] Similarly, at the 2008 Kentucky Derby, the heavy favorite and eventual winner, Big Brown, went to the post with five-to-two odds in his race against nineteen other horses.[89]

In these sporting venues, the reason that the prior probability of a particular team or horse winning is not fifty-fifty is obvious. The choice is not "Either A will win or A will not win." Rather, the choice is "Either A will win, or B will win, or C will win, etc." Similarly, in the paternity context, if all we have is an accusation, the choice is not "Either A is the father or A is not the father," but "Either A is the father, or B is the father, or C is the father, etc." In order for fifty-fifty to represent the actual prior odds of paternity, there would have to be exactly one person other than the suspect who had sexual relations with the mother during the appropriate time span. To reject the use of the fifty-fifty prior probability is not to say that there are endless groups of possible fathers in every case. Rather, it is to assert the simple proposition that the automatic use of the fifty-fifty prior probability is inappropriate because, without knowing other evidence, it is impossible to know how many possible fathers there are.[90]

---

88. Capperspicks.com, March Madness NCAA Tournament Betting Odds Available, http://www.capperspicks.com/forums/online-sportsbook-casino-horse-racing-poker-industry-news/1515-march-madness-ncaa-tournament-betting-odds-available.html (last visited Feb. 5, 2010).

89. Mark Blaudschun, *Triumph, Tragedy at Derby*, BOSTON GLOBE, May 4, 2008, at C1.

90. Thus, the problem with the fifty-fifty probability is not that it operates "upon the assumption 'that the mother and putative father have engaged in sexual intercourse at least once during the period of possible conception.'" State v. Hartman, 426 N.W.2d 320, 326 (Wis. 1988), (quoting *In re* Paternity of M.J.B., 425 N.W.2d 404, 409 (Wis. 1988)), *rev'g* 412 N.W.2d 901 (Wis. App. 1987). Rather, as the New Jersey Supreme Court noted in *Spann*, fifty-fifty odds

> are wholly consistent with a fact pattern that one and only one man had access to and intercourse with the victim and that one of two, and only two, men, including defendant,

Another problem with the seemingly neutral assumption of fifty-fifty prior probability is that it can easily lead to ridiculous, counter-factual results. In *Snake Eyes*,[91] for example, if we used the fifty-fifty prior probability, we would conclude that Al, who was not in the house when the shooting occurred, would be deemed to have the same likelihood of committing the crime as Dennis. This leads to two perverse results: (a) two different people are each given a probability of more than eighty percent of being the only shooter;[92] and (b) someone who is definitely innocent is perceived as far more likely to be guilty than innocent.

Lest one think this is a fanciful case, consider the plight of Donald Cole.[93] A North Carolina district court judge found him to be the biological father of Jonathan Cole, based on evidence showing that the probability of paternity was 95.98%.[94] The judge found that the numerical value for the probability of paternity was more probative than the fact that the purported father had had a vasectomy before Jonathan was born, and that tests showed a sperm count of zero both before and after the birth.[95] The finding of paternity was reversed on appeal, but the lure of an easy number has convinced all-too-many others to opt for counter-factual certainty.

Even while conceding that the assumption of a fifty-fifty prior probability "will not correspond to the facts in most cases of disputed paternity," the *Joint AMA-ABA Guidelines* recommended use of the fifty-fifty assumption as a "useful working hypothesis."[96] The Uniform Parentage Act similarly creates a rebuttal presumption of paternity when there is a probability of paternity of ninety-nine percent, "using a prior probability of 0.50."[97] Many states have specifically adopted this language of the Uniform Parentage Act, specifying use of "a prior probability of 0.50."[98] Among

---

could possibly have been that one man, neither one more likely than the other to be the father.

State v. Spann, 617 A.2d 247, 253 (N.J. 1993); *see also Griffith*, 976 S.W.2d at 248 ("Logically, the prior probability assumes intercourse *could* have occurred and thus the putative father could be the actual father, but the statistic does not necessarily assume intercourse *did* occur.").

91. *See supra* text accompanying note 40.

92. *See supra* note 46 and accompanying text.

93. Cole v. Cole, 328 S.E.2d 446 (N.C. Ct. App. 1985), *aff'd*, 335 S.E.2d 897 (N.C. 1985).

94. *Id.* at 448.

95. *Id.* at 449.

96. *See* Jack P. Abbott, Kenneth W. Sell & Harry D. Krause, *Joint AMA-ABA Guidelines: Present Status of Serologic Testing in Problems of Disputed Parentage*, 10 FAM. L.Q. 247, 262 (1976).

97. UNIF. PARENTAGE ACT § 505(a)(1) (2000) (amended 2002). The presumption also requires "a combined paternity index of at least 100 to 1." *Id.* § 505(a)(2). Initially promulgated in 1973 by the National Conference of Commissioners on Uniform State Laws, the Uniform Parentage Act in its current form is available at http://www.law.upenn.edu/bll/archives/ulc/upa/final2002.pdf.

98. Examples include: ALA. CODE § 26-17-505 (2009); CAL. FAM. CODE § 7555 (West 2004); DEL. CODE ANN. tit. 13, § 8-505 (2009); MINN. STAT. § 257.62 (2009); MO. REV. STAT. § 210.822 (2009); N.Y. FAM. CT. ACT § 532 (McKinney 2009), *unconstitutional as applied by In re* Adoption

those states whose laws do not stipulate the fifty-fifty prior probability, virtually every state still allows its use in creating a probability of paternity.[99]

Some have proposed that, rather than rely on a standard fifty-fifty prior probability, a more accurate assessment can be made by having jurors determine the prior probability for themselves. To assist with the calculation, the jurors would be given a range of different prior probabilities and the probability of paternity associated with each.[100] One court stated,"[t]he expert should present calculations based on assumed prior probabilities of 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 percent."[101]

This proposal, though well-meaning, is hopelessly misguided. On the most basic level, jurors are being asked to quantify the strengths of their subjective opinions. The problem is that people generally cannot condense their thoughts, feelings, and intuitions into a solid number.[102] As Professor J.H. Wigmore wrote, "no one has yet invented or discovered a mode of measurement for the intensity of human belief."[103]

Even if opinions could be turned into numerical probabilities, justice would not be served. A jury will have great difficulty balancing hard numbers "against such fuzzy imponderables as the risk of frame-up or of misobservation, if indeed it is not induced to ignore those imponderables altogether."[104] The danger is that the attempt to concretize what is inherently a subjective analysis will tend to "shift the focus away from such elements as volition, knowledge, and intent, and toward such elements as identity and occurrence—for the same reason that the hard variables tend to swamp the soft."[105]

---

of Sebastian, 879 N.Y.S.2d 677 (2009); N.D. CENT. CODE § 14-20-29 (2009); OKLA. STAT. tit. 10, § 7700-505 (2009); TEX. FAM. CODE ANN. § 160.505 (Vernon 2009); UTAH CODE ANN. § 78B-15-505 (2009); WASH. REV. CODE § 26.26.420 (2010); and WYO. STAT. ANN. § 14-2-705 (2009).

99. *See* George Maha, *Analysis of Genetic Test Results for Courtroom Use*, *in* DISPUTED PATERNITY PROCEEDINGS § 15.08 (Carl W. Gilmore ed., 2008).

100. *See, e.g.*, Michael O. Finkelstein & William B. Fairley, *A Bayesian Approach to Identification Evidence*, 83 HARV. L. REV. 489 (1970). This was the approach used by the court in *State v. Spann*, 617 A.2d 247, 264–65 (N.J. 1993).

101. Plemel v. Walter, 735 P.2d 1209, 1219 (Or. 1987).

102. *See* Ira Mark Ellman & David Kaye, *Probabilities and Proof: Can HLA and Blood Group Testing Prove Paternity?* 54 N.Y.U. L. REV. 1131, 1153 (1979) ("[I]nstructing the jury to follow the chart may be asking it to do something it cannot: to translate a subjective opinion about the non-test evidence into a single probability figure.").

103. 9 J.H. WIGMORE, EVIDENCE § 2497, at 325 (3d ed. 1940).

104. Laurence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329, 1365 (1971).

105. *Id.* at 1366; *see also* Charles R. Nesson, *Reasonable Doubt and Permissive Inferences: The*

One court has rejected having jurors determine their own estimate of prior probability because the resulting trial would be "unduly complicated."[106] Unfortunately, this same court barred the admission of the factually-based "probability of exclusion" because jurors were "apt to confuse" it with the simplified but inaccurate "likelihood of paternity."[107] Thus, in the name of simplicity, the court permitted the use of the inaccurate fifty-fifty prior probability.

This may explain a large part of the reason the legal community has so embraced the unfortunate use of fifty-fifty prior probability. In addition to its seeming facial "neutrality," its use makes the intimidating math of Bayes' Theorem easier to grasp and "more understandable."[108]

To insist upon using the ersatz fifty-fifty probability because it is "more understandable," though, presents "the absurdity of looking for the lost coin under the lamppost solely because the light is better."[109] In the words of Oliver Wendell Holmes, the appeal of the precision of numbers is that they "flatter that longing for certainty and for repose which is in every human mind. But certainty generally is illusion, and repose is not the destiny of man."[110] It has long been understood by real mathematicians that "[f]ar better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question . . . ."[111]

There is no disagreement about the mathematical accuracy of Bayes' Theorem, the statistical meaning of the PI, or the fact that the generalized use of the fifty-fifty prior probability is not necessitated by either science or mathematics. However, it is a task for the legal system to determine whether the use of a fixed fifty-fifty prior probability in order to determine the likelihood of paternity provides an exact answer to the wrong question. It

---

Value of Complexity, 92 HARV. L. REV. 1187, 1225 (1979) ("[A]ny conceptualization of reasonable doubt in probabilistic form is inconsistent with the functional role the concept is designed to play.").

106. Commonwealth v. Beausoleil, 490 N.E.2d 788, 797 n.19 (Mass. 1986). Among the proposals rejected were presenting a chart to jurors showing the effect the blood test results would have on a juror's own estimates of the prior odds of paternity, and providing jurors with a formula so that they could see how their own estimate of prior probability would change the probability of paternity calculation based on an assumed fifty percent prior probability. *Id.*; *see also* Ellman & Kaye, *supra* note 102, at 1152–58; Peterson, *supra* note 60, at 686–89.

107. *Beausoleil*, 490 N.E.2d at 795; *see also* Kofford v. Flora, 744 P.2d 1343, 1351 (Utah 1987) (stating that because of the possibility of confusion, the probability of exclusion should only be admitted when it is in an "extreme range" (quoting Imms v. Clarke, 654 S.W.2d 281, 285 (Mo. Ct. App. 1983))).

108. Butcher v. Kentucky, 96 S.W.3d 3, 7, 9 (2002).

109. William M. Sage, *Judicial Opinions Involving Health Insurance Coverage: Trompe L'Oeil or Window on the World?*, 31 IND. L. REV. 49, 50 (1998). For a related lamppost metaphor, see Allen & Pardo, *supra* note 48, at 119 (stating that a similar analysis was "reminiscent of relying on the lamppost more for support than illumination").

110. Oliver Wendell Holmes, *The Path of the Law*, 10 HARV. L. REV. 457, 466 (1897).

111. John W. Tukey, *The Future of Data Analysis*, 33 ANNALS MATHEMATICAL STAT. 1, 13–14 (1962).

does not matter that "within the relevant community of blood testers, the paternity probability calculations . . . were based upon scientific methods, accepted world-wide, which incorporate both Bayes' Theorem *and* the .5 prior probability."[112] The real issue is whether the use of an unsubstantiated, counterfactual prior probability is relevant for deciding whether a particular suspect is the father. Simply put, "[w]hat is and is not relevant is not appropriately decided by scientists and statisticians."[113] It is "for the trial court, not the scientific community, to determine the relevance of the technique."[114] If the numbers are not accurate, they are irrelevant and should not be used.

But sometimes the numbers are accurate and still should not be used. Such a situation arises with another DNA issue, this time the use of race in describing genetic statistics.

III. RACIALIZED NUMBERS

Imagine that at Dennis's trial for the *Snake Eyes* murder, the prosecution attempted to present evidence that, according to the Bureau of Justice Statistics, homicides are committed by one in 28,574 whites and one in 3,773 blacks.[115] Among the many objections to the admissibility of this evidence would be that there is no reason to believe that race is relevant to the question of Dennis's guilt and that bringing race unnecessarily before the jury would imply that Dennis's race was somehow relevant.

It is astonishing, therefore, that such "race talk" is a commonplace occurrence at criminal trials all over America. Even in cases where there is

112. Kammer v. Young, 535 A.2d 936, 941 (Md. Ct. Spec. App. 1988); *see also* Brown v. Smith, 526 S.E.2d 686, 689 (N.C. Ct. App. 2000) ("Most, if not all, laboratories in the United States use a prior probability of 0.5 in calculating the genetic probability of paternity."); Griffith v. State, 976 S.W.2d 241, 245 n.2 (Tex. App. 1998) ("[N]early a million paternity tests in the U.S. were conducted using DNA or HLA methods, each using the .5 prior probability calculation."); M. v. Marvin S., 656 N.Y.S.2d 802, 805 (N.Y. Fam. Ct. 1997) ("[T]he utilization by a laboratory of the 0.5 figure is a nationally accepted convention and that all the major laboratories use this figure for paternity test reporting purposes.").

113. United States v. Jenkins, 887 A.2d 1013, 1025 (D.C. 2005); *accord* People v. Nelson, 185 P.3d 49, 65 (Cal. 2008), *cert denied*, 129 S.Ct. 357 (2008).

114. *Nelson*, 185 P.3d at 65; *see also Jenkins*, 887 A.2d at 1024 ("This debate does not address the underlying principles, math, or science behind the various formulas. . . . It is a disagreement over relevance."). *See generally* Kaye, *Rounding Up the Usual Suspects*, *supra* note 31, at 448 (referring to a "question . . . of logical relevance" as opposed to one of "general acceptance" or "scientific validity").

115. These numbers are derived from BUREAU OF JUSTICE STATISTICS, U.S. DEP'T OF JUSTICE, HOMICIDE TRENDS IN THE UNITED STATES, http://bjs.ojp.usdoj.gov/content/pub/pdf/htius.pdf (last visited Feb. 6, 2010).

no evidence of the perpetrator's race, jurors are often presented information in explicitly racial terms, describing a DNA match between the defendant and genetic material found at the crime scene. For example, in a typical case, the prosecution's expert testified that "[d]efendant's genetic profile would be expected to occur in one of 96 billion Caucasians, one of 180 billion Hispanics, and one of 340 billion African-Americans."[116] Indeed, when the race and ethnicity of the perpetrator is unknown, "providing statistics from several racial groups is the standard way of assessing the significance of a match . . . ."[117]

Those not well-versed in thinking about statistics can easily be overwhelmed when told that scientists use racial categories to create such extraordinarily intimidating numbers as "one of 340 billion African-Americans." The impressiveness of those numbers, however, cannot be permitted to prevent the legal system from making its own value judgment about the significance of the fact that there are no actual definitions to delineate the racial categories. The numbers should also not prevent judges from recognizing the harm that results from permitting courtroom discussions of race when race would otherwise be irrelevant.

Scientifically, the DNA analysis for a criminal match is similar, but not identical, to that for paternity matching. In the latter, we are trying to see if a suspect's DNA is consistent with that of whoever contributed half of the child's DNA. In criminal matching, we are trying to determine whether two DNA samples are identical.

When most people think about DNA, they focus on the many thousands of genes, which are linked segments of DNA. Genes have specific functions that determine all the physical traits that we inherit from our biological parents.[118] Genes, however, make up only a tiny percentage of our DNA.[119] The vast majority of human DNA, estimated at ninety-seven percent, is known as "non-coding" material or "junk DNA," because it serves no known function.[120]

---

116. People v. Wilson, 136 P.3d 864, 867 (Cal. 2006); *see also, e.g.*, State v. Spann, 617 A.2d 247, 251 (N.J. 1993) ("The State's expert stated that the blood and tissue samples, combined with statistical data reflecting the number of men with the relevant genes, excluded 99% of the North American black male population as possible fathers.").

117. D. H. Kaye, *Logical Relevance: Problems with the Reference Population and DNA Mixtures in* People v. Pizarro, 3 L. PROBABILITY & RISK 211, 214 (2004).

118. Usually, important traits are the product of the relationship between many different genes. Kaye & Sensabaugh, *Reference Guide on DNA Evidence*, *supra* note 51, at 491.

119. NAT'L COMM. ON THE FUTURE OF DNA EVIDENCE, U.S. DEP'T OF JUSTICE, THE FUTURE OF FORENSIC DNA TESTING: PREDICTIONS OF THE RESEARCH AND DEVELOPMENT WORKING GROUP 12 (2000) [hereinafter FUTURE OF FORENSIC DNA TESTING], *available at* http://www.ncjrs.gov/pdffiles1/nij/183697.pdf.

120. *Id.*

About 99.9% of DNA is identical between any two individuals.[121] Differences in either genes or junk DNA are identified as "alleles."[122] A position on a specific chromosome, called a "locus," where almost all humans have the same DNA sequence, is termed "monomorphic."[123] A locus with multiple possible alleles is termed "polymorphic."[124] The more variations there are among alleles, the easier it is to make distinctions between DNA samples.[125] Because junk DNA tends to be highly polymorphic, that is, it contains far greater variation among individuals, it is used for forensic identification.[126]

What that means is that when an expert testifies about the likelihood of DNA appearing in different racial groups, she is only referring to "non-coding," or junk, DNA. Thus, she is not reporting on the DNA that determines skin color or any other physical or biological characteristics associated with specific races; the common assumption to the contrary is completely, if not dangerously, misplaced.

While no single gene or collection of genetic material is specifically associated with any one race, geneticists have determined that some non-coded material is found in greater frequency in some population groups than others. It is critical to recognize that not everyone in a particular ethnic or racial group will have that same genetic material. Moreover, those in other population groups may very well share that particular DNA. Thus, when expert testimony is given about race and DNA, the expert is essentially "making highly probabilistic statements about suspects and the likely ethnic, racial, or cultural populations from which they can be identified—statistically."[127]

---

121. Kaye & Sensabaugh, *Reference Guide on DNA Evidence*, *supra* note 51, at 491.

122. *Id.* at 492.

123. *Id.* at 492, 571.

124. *Id.* at 492.

125. *See id.* at 493.

126. *See id.* at 492 n.25; *see also* DNA Analysis Backlog Elimination Act of 2000, H.R. REP. NO. 106-900(I),pt. 1, at 27 (2000), Pub. L. No.106-546, 2000 U.S.C.C.A.N. 2726 (codified at 42 U.S.C. § 13701) (stating that for privacy reasons, the non-coded regions "were purposely selected because they are not associated with any known physical or medical characteristics").

127. Troy Duster, *Selective Arrests, an Ever-Expanding DNA Forensic Database, and the Specter of an Early-Twenty-First-Century Equivalent of Phrenology*, *in* DNA AND THE CRIMINAL JUSTICE SYSTEM: THE TECHNOLOGY OF JUSTICE 314, 325 (David Lazer ed., 2004); *see also* State v. Spann, 617 A.2d 247, 251 n.3 (N.J. 1993) ("Since the incidence of different blood groups, as well as HLA types, varies with race and to a lesser extent geography, gene-frequency tables are derived from population studies of different racial groups.").

Because of this statistical variation, "[t]he FBI's databases are divided along racial lines."[128] The FBI has divided its national DNA database into five separate population groups: African-Americans, United States Caucasians, Hispanics, Far East Asians, and Native Americans.[129] It is from these FBI databases that courtroom experts derive the racial genetic probabilities that they proclaim. But these probabilistic statements, especially when presented with the mathematical certainty of "one of 340 billion African-Americans," mask a series of problems that can escape those easily blinded by numbers.

First, there is the question of defining racial categories. If we are to divide Americans in a scientific fashion into five population groups, we need to have a working definition for each group. Obviously, if we are to distinguish apples from oranges, we need to know the difference between apples and oranges.

The demarcation between racial categories is especially important for the national DNA database. The federal Combined DNA Identification System (CODIS) is a three-tiered system.[130] Local law enforcement agencies collect the DNA data from those they arrest and create a Local DNA Index System (LDIS).[131] Each state then combines the local profiles into a State DNA Index System (SDIS).[132] Each of these state compilations is then combined with the FBI's database into a National DNA Index (NDIS),[133] which contains more than seven million "offender profiles."[134]

Were only one entity to be charged with compiling and categorizing individuals by race, a clear definition of the categories would be necessary. But when thousands of individual local law enforcement agencies are deciding who is "African-American," who is "Caucasian," and who is "Hispanic," such clarity is essential.

---

128. People v. Dalcollo, 669 N.E.2d 378, 381 (Ill. App. Ct. 1996).

129. *See* JOHN M. BUTLER, FORENSIC DNA TYPING: BIOLOGY, TECHNOLOGY, AND GENETICS OF STR MARKERS 282–83 (2d ed. 2005); *cf.* Bruce Budowle et al., *CODIS STR Loci Data from 41 Sample Populations*, 46 J. FORENSIC SCI. 453, 453 (2001) (stating that in the United States, for purposes of DNA analysis, African-American, U.S. Caucasian, Hispanics, Far East Asians, and Native Americans make up the five "major population groups"). *See generally* Jonathan Kahn, *Race, Genes, and Justice: A Call to Reform the Presentation of Forensic DNA Evidence in Criminal Trials* (2008), *available at* http://works.bepress.com/jonathan_kahn/1.

130. Aaron P. Stevens, Note, *Arresting Crime: Expanding the Scope of DNA Databases in America*, 79 TEX. L. REV. 921, 927 (2001).

131. *Id.*

132. *Id.* at 927–28.

133. *Id.* at 928.

134. Federal Bureau of Investigation, Laboratory Services, CODIS—NDIS Statistics, http://www.fbi.gov/hq/lab/codis/clickmap.htm (last visited Feb. 6, 2010); *see also* Erin Murphy, *The New Forensics: Criminal Justice, False Certainty, and the Second Generation of Scientific Evidence*, 95 CAL. L. REV. 721, 739–40 (2007). The NDIS consists primarily of those who are charged with, or have been convicted of, serious crimes including felonies and other crimes of violence. 42 U.S.C. §§ 14132, 14135a(d) (2006).

Incredibly, there are no definitions of the particular racial categories that are commonly used in courts throughout the nation. An FBI-funded report by the National Research Council not only recommended the use of racial categories[135] but also admitted that there could be no uniform way of delineating the categories:

> There is no generally agreed-on vocabulary for treating human diversity. Major groups are sometimes designated as races, and at other times as ethnic groups. Ethnic group is also used to designate subgroups of major groups.... [G]roups are mixed, all the classifications are fuzzy at the borders, and the criteria for membership are variable. For such reasons, some assert that the word race is meaningless. *But the word is commonly used and generally understood, and we need a vocabulary.*[136]

In other words, despite the authoritative sound of the race-based genetic statistical evidence, the actual classification system is no more precise, consistent, or objective than Justice Stewart's notorious description of obscenity: "*I know it when I see it.*"[137] The way that race is "generally understood" is entirely subjective and non-scientific, based on outward appearance and the societal association of that appearance with a particular racial label: "Even though we may feel confident of our visual perceptions and racial or ethnic conclusions, we know that this kind of classification is dismally inaccurate."[138]

One difficulty with acknowledging this inaccuracy is that race is in our nation's DNA. From the initial racialization of slavery, racial definitions have been part of our national discourse.[139] But the actual placement of

---

135. Paul C. Giannelli, *Forensic Science: Under the Microscope*, 34 OHIO N.U. L. REV. 315, 327 n.86 (2008).

136. NRC II, *supra* note 51, at 57–58 (emphasis added). This report was a follow-up to an earlier FBI-funded report, NATIONAL RESEARCH COUNCIL, DNA TECHNOLOGY IN FORENSIC SCIENCE (1992).

137. Jacobellis v. Ohio, 378 U.S. 184, 197 (1964) (Stewart, J., concurring) (emphasis added). The full quote referred to the difficulty in defining what he termed "hard-core pornography":

> I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description; and perhaps I could never succeed in intelligibly doing so. But I know it when I see it, and the motion picture involved in this case is not that.

138. Mildred K. Cho, *Racial and Ethnic Categories in Biomedical Research: There Is No Baby in the Bathwater*, 34 J.L. MED. & ETHICS 497, 498 (2006).

139. *See, e.g.*, Audrey Smedley & Brian D. Smedley, *Race as Biology Is Fiction, Racism as a Social Problem Is Real: Anthropological and Historical Perspectives on the Social Construction of Race*, 60 AM. PSYCHOLOGIST 16, 22 (2005) ("The ideology [of race] arose as a rationalization and

individuals in racial categories, even if "generally understood," has always been imprecise. Ironically, at the same time the Supreme Court in *Plessy v. Ferguson* was upholding segregation as the inevitable result of "racial instincts," the Court was also acknowledging that there was no agreed-upon distinction between races: "[T]he proportion of colored blood necessary to constitute a colored person, as distinguished from a white person, is one upon which there is a difference of opinion in the different States."[140]

Today there is still no consensus on the meaning of race. When the Oxford English Dictionary attempted to define "race," it recognized that "the term is often used imprecisely; even among anthropologists there is no generally accepted classification or terminology."[141] Moreover, despite our common usage, racial categories are not "discrete."[142] There is no bright line separating the categories. Unlike with fish or fowl, racial categories are not "mutually exclusive."[143]

Still, if, as the National Research Council noted, race is "commonly used," how are we able to categorize individuals by race? Many have argued that race is "socially constructed," meaning that race is not innate and unchanging.[144] People look at the external physical traits, especially skin color, and associate them with a particular race. These categories, however, are "socially fluid": "For example, in the US, people with ancestry from India are sometimes labeled Asian and sometimes labeled white or 'Caucasian'; they are not classified in the same way in the UK as in the US."[145]

---

justification for human slavery at a time when Western European societies were embracing philosophies promoting individual and human rights, liberty, democracy, justice, brotherhood, and equality.").

140. Plessy v. Ferguson, 163 U.S. 537, 552 (1896). Homer Plessy was described as "seven eighths Caucasian and one eighth African blood." *Id.* at 541. *See* Saint Francis College v. Al-Khazraji, 481 U.S. 604, 610–11 (1987) ("In the middle years of the 19th century, dictionaries commonly referred to race as a 'continued series of descendants from a parent who is called the stock,' 'the lineage of a family,' or 'descendants of a common ancestor,' [and] [i]t was not until the 20th century that dictionaries began referring to the Caucasian, Mongolian, and Negro races, or to race as involving divisions of mankind based upon different physical characteristics.") (citations omitted).

141. OXFORD ENGLISH DICTIONARY 69 (2d ed. 1989).

142. Martha Chamallas, *Questioning the Use of Race-Specific and Gender-Specific Economic Data in Tort Litigation: A Constitutional Argument*, 63 FORDHAM L. REV. 73, 113 (1994) [hereinafter Chamallas, *Questioning the Use*] ("In a multi-racial society, such as the United States, people do not fall naturally into discrete racial groupings.").

143. Duster, *supra* note 127, at 325.

144. Erik Lillquist & Charles A. Sullivan, *The Law and Genetics of Racial Profiling in Medicine*, 39 HARV. C.R.-C.L. L. REV. 391, 394 (2004).

145. Mildred K. Cho & Pamela Sankar, *Forensic Genetics and Ethical, Legal, and Social Implications Beyond the Clinic*, 36 NATURE GENETICS SUPPLEMENT at S8, S9 (2004), *available at* http://backintyme.com/admixture/cho01.pdf.

Not only will racial labels vary by geography, they can also vary by time. It is not at all clear that Homer Plessy, who was generally considered to "look White" would be considered African-American today. Indeed, people may even change how they self-identify, altering the racial group to which they say they belong.[146]

The "incoherence" of race as a category is not contradicted by the fact that there is some statistical correlation between the frequency of certain alleles and our ill-defined racial categories.[147] What the DNA variations actually are correlated to is, at best, a partial ancestral geographic origin.[148] In other words, the DNA statistics might signal where some of one's ancestors originated.

But even to the extent that geography is in our genes, the five racial categories of CODIS cannot capture the reality of America. Unlike ancient, insular societies, Americans do not stay isolated in neatly-definable groupings: "After hundreds of years of sexual mixings, there continues to be 'no socially sanctioned in-between classification' of 'race' in America."[149] Accordingly, the correlation of DNA and geography confirms the incoherence of the CODIS statistical analysis, "the reality being that the diversity of human biology has little in common with socially constructed 'racial' categories."[150]

Thus, the introduction of racially-based DNA numbers into a courtroom proceeding is fundamentally misleading. The geographic origins of a particular long-dead ancestor, which might be conveyed by those numbers, is simply not the same as the social classification a juror may associate with a particular outward physical appearance. When jurors hear that a particular combination of alleles occurs "in one of 96 billion Caucasians, one of 180 billion Hispanics, and one of 340 billion African-Americans," they will assume the number applies to the racial category in which they have placed

146. *See id.* ("[I]ndividual self-classification is not stable; for example, one US study found that one-third of people change their own self-identified race or ethnicity in two consecutive years."); *see also* Christopher J. L. Murray et al., *Eight Americas: Investigating Mortality Disparities Across Races, Counties, and Race-Counties in the United States*, 3 PLOS MED. 1513, 1521 (2006) ("The most important limitation of the data used for our analysis is that reported race in the census, used for population estimates, may be different from race in mortality statistics, where race may be reported by the family, the certifying physician, or the funeral director.").

147. Sharona Hoffman, *Is There a Place for "Race" as a Legal Concept?*, 36 ARIZ. ST. L.J. 1093, 1096 (2004).

148. *See generally* Duana Fullwiley, *The Biologistical Construction of Race: 'Admixture' Technology and the New Genetic Medicine*, 38 SOC. STUD. SCI. 695 (2008).

149. McMillan v. City of New York, 253 F.R.D. 247, 251 (E.D.N.Y. 2008) (quoting Smedley & Smedley, *supra* note 139, at 20).

150. *McMillan*, 253 F.R.D. at 250 (quoting Smedley & Smedley, *supra* note 139, at 20).

the defendant based on a subjective interpretation of outward physical appearances, regardless of the defendant's actual (and unknown) ancestral origins. The influence of the numbers masks the unspoken assumptions.

But there is a greater problem with the legal profession's awe of numbers. By relinquishing authority to those who control the numbers, courts have abandoned their responsibility to consider the harm caused by unnecessary "race speech" in court. This cavalier attitude was expressed by the California Supreme Court when it endorsed the admission of expert testimony presenting a range of racially-characterized genetic profile frequencies: "Presenting the objective data in the manner in which such information is collected and analyzed within the scientific community does not inject inappropriate racial assumptions or issues into the litigation."[151]

Courts must not be so intimidated by "objective data" that they fail to consider the harm created whenever race is introduced into a courtroom discussion. In the story of *Snake Eyes*,[152] the race of Dennis was irrelevant; indeed, I suspect, it was outside of most readers' thoughts until racial testimony brought the issue to the reader's attention. Courts should not casually permit the insertion of race into a juror's analysis.

When the prosecutor puts forth racially-categorized statistics, it immediately raises the question of the relevance of race. Imagine if, at a trial of an African-American defendant, a juror is told that the DNA found at the scene of the crime matched the defendant and that this type of DNA occurred in one out of 10,000 United States Caucasians and one out of 10 billion African Americans. The relevance of the DNA would depend on the jury's determination as to how likely it was that the crime had been committed by an African-American. This is so because there would be thousands of Caucasians whose DNA would match that of the DNA found at the crime scene, but very few, if any, other African-Americans' DNA would match. Thus, if the jury believed that an African-American committed, or would be likely to commit, the crime, the overwhelming likelihood would be that the defendant was the culprit.

This sort of racial thinking is inherent in the use of racial categories because "by highlighting, without compelling justification, the racial distinctions that have historically divided us," the government is expressing "an improperly *divisive* conception of the public."[153] Historically, and unfortunately even today, "[t]he word 'race' suggests that human beings can

---

151. People v. Wilson, 136 P.3d 864, 871 (Cal. 2006); *see also* Edward J. Imwinkelried & D.H. Kaye, *DNA Typing: Emerging or Neglected Issues*, 76 WASH. L. REV. 413, 449 (2001) ("No group is singled out for special treatment, and no one is penalized because of hostility toward race.").

152. *See supra* text accompanying note 40.

153. Elizabeth S. Anderson & Richard H. Pildes, *Expressive Theories of Law: A General Restatement*, 148 U. PA. L. REV. 1503, 1538 (2000).

be divided into subspecies, some of which are morally, intellectually, and biologically inferior to others."[154]

The Supreme Court has recognized this danger, even when no particular group was being treated "differently." In *Anderson v. Martin*,[155] the Court struck down a Louisiana statute requiring that ballots designate the race of candidates for elective office. The Court stated that, although Louisiana was not restricting any voter's individual choice, "by directing the citizen's attention to the single consideration of race or color, the State indicates that a candidate's race or color is an important—perhaps paramount—consideration in the citizen's choice . . . ."[156] The unconstitutional evil arose because, by "placing a racial label on a candidate[,] . . . the State furnishes a vehicle by which racial prejudice may be so aroused as to operate against one group because of race and for another."[157]

When not under the hypnotic influence of numbers, courts readily recognize the danger of governmental use of race-based categorizations: "[R]acial classifications are simply too pernicious to permit any but the most exact connection between justification and classification."[158] Accordingly, the Supreme Court has frequently declared that, "all racial classifications, imposed by whatever federal, state, or local governmental actor, must be analyzed by a reviewing court under strict scrutiny."[159] A racial classification will only pass strict scrutiny if it is necessary for furthering some compelling interest and is narrowly tailored to further that interest.[160]

Conceding that the accurate determination of a criminal defendant's guilt or innocence is compelling still leaves the critical question of whether

---

154. Hoffman, *supra* note 147, at 1099.

155. 375 U.S. 399 (1964). The Louisiana law stated:

> Every application for or notification or declaration of candidacy, and every certificate of nomination and every nomination paper filed in any state or local primary, general or special election for any elective office in this state shall show for each candidate named therein, whether such candidate is of the Caucasian race, the Negro race or other specified race.

Id. at 400 n.1. The law also required that "[t]he racial designation on the ballots shall be in print of the same size as the print in the names of the candidates on the ballots." *Id.*; *see also* LA. REV. STAT. ANN. §§ 18:1174.1(A), (C) (Supp. 1960).

156. *Anderson*, 375 U.S. at 402.

157. *Id.*

158. Gratz v. Bollinger, 539 U.S. 244, 270 (2003) (quoting Fullilove v. Klutznick, 448 U.S. 448, 537 (1980) (Stevens, J., dissenting)).

159. Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 227 (1995); *accord* Johnson v. California, 543 U.S. 499, 505 (2005).

160. *See* Grutter v. Bollinger, 539 U.S. 306, 327 (2003) ("When race-based action is necessary to further a compelling governmental interest, such action does not violate the constitutional guarantee of equal protection so long as the narrow-tailoring requirement is also satisfied.").

the racial categorization of DNA is "necessary" for and "narrowly tailored" to the making of that determination.[161] Obviously, if there were any racially neutral way of presenting the DNA information in a meaningful fashion, the Constitution would require the FBI and courts to forego the race-based approach.

Fortunately, there is a relatively simple mathematical solution to this problem that will fulfill the "desire for a race-blind figure in a general-population case."[162] All that is required is the use of what mathematicians call a "corrective factor."[163] Mathematicians often add so-called corrective factors to their equations so that their generalized theoretical predictions can more accurately reflect particular factual situations.[164] For example, when calculating back pay in one case, the Equal Employment Opportunity Commission determined that using only the hourly wage would not represent the full amount of money lost, and proposed "a corrective factor to be placed in the formula which would accurately reflect the effect of overtime hours."[165] In a similar fashion, the National Committee on the Future of DNA Evidence has shown that, by placing the appropriate "corrective factor" in the equations for calculating genetic probabilities, "the necessity for group classification could be avoided by using an overall U.S. database."[166] Significantly, the Committee reported that it was able to

---

161. One instance in which such racial categorization might be necessary is when dealing with ethnic subpopulations. For example, assume that there is a small group whose members share a genetic anomaly that is not seen with any other group. Using a general population database might lead to a finding that the defendant was probably guilty (since very few Americans match the DNA found at the crime scene). However anyone in his subgroup would have matched that DNA as well. In such a case, use of ethnic data bases would be appropriate. *See* Kaye & Sensabaugh, *Reference Guide on DNA Evidence*, *supra* note 51, at 526; *see also* R. C. Lewontin, Letter to the Editor, *Which Population?*, 52 AM. J. HUM. GENETICS 205, 205 (1993).

162. David H. Kaye, *DNA Probabilities in* People v. Prince*: When Are Racial and Ethnic Statistics Relevant?*, *in* PROBABILITY AND STATISTICS: ESSAYS IN HONOR OF DAVID A. FREEDMAN 289, 300 (Deborah Nolan & Terry Speed eds., 2008), *available at* http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.imsc/1207580069.

163. *See, e.g.*, ELLERY WILLIAMS DAVIS & WILLIAM CHARLES BRENKE, THE CALCULUS 170 (1912) (stating that, for a particular formula used to find the measure of the bending of a curve, "the corrective factor . . . gives a better measure of the bending . . . .").

164. Andre A. Moenssens, *Handwriting Identification Evidence in the Post-*Daubert *World*, 66 UMKC L. REV. 251, 277, 287 n.154 (1997).

165. Bowe v. Colgate-Palmolive Co., 443 F. Supp. 696, 711–12 (S.D. Ind. 1977).

166. FUTURE OF FORENSIC DNA TESTING, *supra* note 120, at 5. The specific corrective factor they gave is $\theta = .03$ for the formula used to determine match probabilities for various alleles (designated "i") and the frequency (designated "p_i") in which they occur:

$$[2\theta + (1-\theta)\,p_i][3\theta + (1-\theta)\,p_i] / [(1 + \theta)(1 + 2\theta)].$$

Id. at 24.

convert to a race-neutral formula so easily because genetic differences "are mainly between individuals rather than between group averages."[167]

Because a scientifically sound basis exists for using general database probabilities, the continued use in criminal trials of race-based statistics cannot be legally justified. Judges must tell genetic statisticians that the emphasis on race "makes racial divisions too salient."[168] The legal system must reassert itself and prohibit "the placing of the power of the State behind a racial classification that induces racial prejudice . . . ."[169]

IV. BIGOTED NUMBERS

Sometimes, the problem with numbers is not that they induce prejudice in others, but that the very use of numbers is unknowingly bigoted. Consider the case of an insurance company which, in 1962, charged African-Americans more than twenty-eight percent more than Caucasians for the identical life insurance policy.[170] The insurance company defended this discriminatory treatment as justified by statistics showing that, at the time of the policy, African-Americans did not, on average, live as long as Caucasians.[171]

Or consider the 2004 case of a three-month-old Native American girl who was killed when her father, in a rage, threw her to the ground, causing her to hit her head against the bathroom sink.[172] After the father was convicted of voluntary manslaughter, the court called for an expert to calculate the lost future income of the infant so that restitution could be awarded for the income she would have earned but for her death at the hands of the defendant.[173] The expert concluded that the restitution should be

167. *Id.* at 5.

168. Anderson & Pildes, *supra* note 153, at 1538.

169. Anderson v. Martin, 375 U.S. 399, 402 (1964).

170. *See In re* Monumental Life Ins. Co., 365 F.3d 408, 416–20 (5th Cir. 2004). The court noted that for a "20 Pay Life" policy, in which the insured pays premiums for twenty years and is then fully insured for the rest of his or her life, a twenty-year-old African-American was charged a weekly premium of $0.41 for a $500 policy, while a twenty-year-old Caucasian was only charged $0.32. *Id.* at 412 n.4.

171. *Id.* at 412 n.2; *see also* Arline T. Geronimus et al., *Inequality in Life Expectancy, Functional Status, and Active Life Expectancy Across Selected Black and White Populations in the United States*, 38 DEMOGRAPHY 227, 235 (2001).

172. United States v. Bedonie, 317 F. Supp. 2d 1285, 1291 (D. Utah 2004), *aff'd sub nom.* United States v. Serawop, 505 F.3d 1112 (10th Cir. 2007).

173. *Id.* at 1292. The restitution award was authorized by the Mandatory Victims Restitution Act, which provides that families of victims of certain crimes will be awarded, as restitution, a sum representing the total income lost due to the crime. 18 U.S.C. § 3663A(b)(2)(C) (2006).

reduced to account for the victim's race and sex; he thus recommended an award of $171,366, which was more than forty-four percent lower than his estimate of lost wages for all Americans of that age, $308,633.[174]

One can make a numeric defense of such disparate treatment. After all, statistics do show that, on average, African-Americans die five years earlier than white Americans.[175] Statistics also show that African-Americans earn, on average, less than whites, and women, on average, earn less than men.[176] Moreover, numbers show that "women have, on average, spent fewer years in the workforce than men, largely because many women have taken time off from work in order to raise children."[177]

Backed by such hard numbers, most courts have been quite willing to make decisions which result in different financial outcomes, depending on the race and gender of the parties. As Professor Martha Chamallas has noted: "[I]t is commonplace for expert witnesses to rely on gender and race-based tables to determine both the number of years that a plaintiff would likely have worked (work/life expectancy) and the likely annual income the plaintiff would have earned."[178] Thus, in reading opinions awarding damages, it is not unusual to read statements such as "in 2003, an African-American female, aged 65, born in and living in the United States, has an additional life expectancy of 18.5 years";[179] "Plaintiff presented evidence from an economics expert . . . as to the demonstrated earning capacity of

174. *Bedonie*, 317 F. Supp. 2d at 1316. The court ultimately rejected the expert's recommendation and awarded the amount calculated without the discount for race and sex. *Id.* at 1322.

175. Based on the age-specific death rates prevailing for the actual population in 2004, the National Center for Health Statistics reported that the average white American lives for 78.3 years, while the average African-American lives for 73.1 years. *See* Elizabeth Arias, *United States Life Tables, 2004*, 56 NAT'L VITAL STAT. REP. 1 (2007), *available at* http://www.cdc.gov/nchs/data/nvsr/nvsr56/nvsr56_09.pdf.

176. U.S. Bureau of Labor Statistics, U.S. Dep't of Labor, *Highlights of Women's Earnings in 2007*, at 8 tbl.1 (2008), *available at* http://www.bls.gov/cps/cpswom2007.pdf. *See also* Caron v. United States, 410 F. Supp. 378, 398 (D.R.I. 1975) ("One does not need expert testimony to conclude that there is inequality in the average earnings of the sexes."), *aff'd*, 548 F.2d 366 (1st Cir. 1976).

177. Childers v. Sec'y of Health & Human Servs., No. 96-194V, 1999 U.S. Claims LEXIS 76, at *56 (Fed. Cl. Mar. 26, 1999); *see also* Sherri R. Lamb, Note, *Toward Gender-Neutral Data for Adjudicating Lost Future Earning Damages: An Evidentiary Perspective*, 72 CHI.-KENT L. REV. 299 (1996) ("[W]orklife tables provide an average for the group, reflecting the historical pattern of actual years worked, incorporating rates of unemployment, both voluntary and involuntary, as well as incorporating an expected retirement age."). There are also disparities in the average worklife for African-Americans as compared to whites: "If minority men have historically been incarcerated at a much higher rate than white men, race-based worklife estimates predict that they will continue to work fewer years than whites." Martha Chamallas, *Civil Rights in Ordinary Tort Cases: Race, Gender, and the Calculation of Economic Loss*, 38 LOY. L.A. L. REV. 1435, 1439 (2005) [hereinafter Chamallas, *Civil Rights in Ordinary Tort Cases*].

178. Chamallas, *Civil Rights in Ordinary Tort Cases*, *supra* note 177, at 1438.

179. Black v. Columbus Pub. Sch., No. 2:96-CV-326, 2007 WL 2713873, at *1 (S.D. Ohio Sept. 17, 2007).

someone of plaintiff's race, sex, age, and educational level";[180] and "future earnings [were calculated] based on the average earnings of a college-educated female of her age."[181]

Even the federal government relies on sex-based tables. In calculating the minimum funding requirements for certain pension plans, the Internal Revenue Service (IRS) provides mortality tables that can be used for determining the current liability for individuals who are entitled to benefits on account of disability. The IRS explains that these mortality tables are "gender-distinct because of significant differences between expected male mortality and expected female mortality."[182] Even more emphatically, the IRS requires that those who wish to use alternate mortality tables for their pension plans must use tables that treat men and women differently: "Separate mortality tables must be established for each gender under the plan."[183]

One might expect that even statistically-justified race and gender distinctions would be met with heightened scrutiny by the courts. Race and gender are, after all, called "suspect classes" because we "suspect" that racial and gender classifications are based on stereotyped views of groups and we "suspect" that, as in the past, these distinctions have the purpose or effect of harming those in the disfavored category.[184] "Yet surprisingly the reported cases have almost completely neglected the question."[185] It is as if,

180. Athridge v. Iglesias, 950 F. Supp. 1187, 1192 (D.D.C. 1996).

181. Forman v. Korean Airlines Co., 84 F.3d 446, 449 (D.C. Cir. 1996); *see also* Gonzalez v. City of Franklin, 383 N.W.2d 907, 913 (Wis. Ct. App. 1986) ("Here, for purposes of determining how many replacement prostheses Gonzalez might need in the future, Gonzalez's counsel used at trial a mortality table which breaks down the populace by race (white and black) and sex (male and female). The figure for white males went to the jury."); Drayton v. Jiffee Chem. Corp., 591 F.2d 352, 368 (6th Cir. 1978) ("We have considered as well her sex, her race, the necessarily limited evidence concerning her mental capabilities, and her psychological makeup."); Feldman v. Allegheny Airlines, 382 F. Supp. 1271, 1286 (D. Conn. 1974) (stating that eight years is the "middle of the range of a professional woman's likely hiatus from her principal occupation in order to raise a family"), *aff'g in part*, 524 F.2d 384 (2nd Cir. 1975); Frankel v. Heym, 466 F.2d 1226, 1229 (3d Cir. 1972) (stating that female plaintiff would probably marry and have children "with consequent substantial interruptions of gainful employment").

182. Mortality Tables for Determining Present Value, 73 Fed. Reg. 44632, 44633 (July 31, 2008) (codified at 26 C.F.R. pt. 1).

183. Mortality Tables Used to Determine Present Value, 26 C.F.R. § 1.430(h)(3)-2(c)(1)(i) (2009).

184. *See, e.g.*, Korematsu v. United States, 323 U.S. 214, 216 (1944) (stating that "all legal restrictions which curtail the civil rights of a single racial group are immediately suspect").

185. United States v. Bedonie, 317 F. Supp. 2d 1285, 1315 (D. Utah 2004), *aff'd sub nom.* United States v. Serawop, 505 F.3d 1112 (10th Cir. 2007). There are a handful of cases in which judges have recognized the dangers posed by race and gender based statistics. *See, e.g.*, McMillan v. City of New York, 253 F.R.D. 247, 256 (E.D.N.Y. 2008); Wheeler Tarpeh-Doe v. United States, 771 F.

when faced with race- or gender-based statistics, "we tend not to notice the discrimination and to accept it as natural and unproblematic."[186]

The power of these numbers can be so great that even well-meaning judges shrink from confronting them. In holding that lost wages for a female plaintiff needed to be reduced to reflect the average woman's lower salary, one judge bemoaned: "I am constrained to agree with the defense that the present value of prospective earnings, female wages before taxes must be used. However sympathetic this Court may be to equality in employment, it must look to the reality of the situation and not be controlled by its own convictions."[187]

The true "reality of the situation," however, is that a reluctance to fully understand what numbers can and cannot tell us has caused the justice system to accept and enforce needless discrimination. The reliance on race- and sex-based statistics should be rejected as both bad mathematics and bad policy.

The first mistake made by those who rely on race- and sex-based statistics is that they ignore one of the cardinal principles of statistics: correlation does not prove causation.[188] The fact that the month with the fewest days has the most snow days does not imply either that short months cause snow or that snow causes short months.

Nonetheless, statistical correlation may still be relevant for predicting the future. We can often use past experience to guess what is likely to happen in the future. Thus, over a span of several years, we can expect that the shortest month generally will continue to be the one that tends to experience the most snow days.

There is a critical assumption, though, which enables us to use statistics of what has already happened to predict what is still to come. For the past to predict the future, the future must resemble the past.[189] To continue our short month-snow day analogy, assume that a new leader takes power and changes the calendar. Declaring that summer vacations are too long and wasteful, this despot decrees that July and August shall henceforth only have twenty-one days, and that the remaining ten months would each get an

---

Supp. 427, 455 (D.D.C. 1991), *rev'd sub nom.* Tarpeh-Doe v. United States, 28 F.333d 120 (D.C. Cir. 1994); Reilly v. United States, 665 F. Supp. 976, 997 (D.R.I. 1987), *aff'd in part*, 863 F.2d 149 (1st Cir. 1988); Hartford Accident & Indem. Co. v. Ins. Comm'r of Pa., 482 A.2d 542, 582 (Pa. 1984).

186. Chamallas, *Civil Rights in Ordinary Tort Cases*, *supra* note 177, at 1442.

187. Caron v. United States, 410 F. Supp. 378, 397–98 (D.R.I. 1975), *aff'd*, 548 F.2d 366 (1st Cir. 1976).

188. *See, e.g.*, SHERRI L. JACKSON, RESEARCH METHODS AND STATISTICS: A CRITICAL THINKING APPROACH 15 (2003) ("Correlation does not imply causation.") (emphasis omitted).

189. *See* Lamb, *supra* note 177, at 329–30 ("Statistical tables 'predict' the future only to the extent that the future resembles the past; a predictor is efficient only if past correlations persist throughout the period in which the predicted event will occur.") (citation omitted).

additional two days. Suddenly, our prediction that the shortest month will have the most snow days is obsolete, even though the statistical analysis on which it was based remains unchanged.

Similarly, a law student in 1963, wondering if his future granddaughter would attend law school, would have been badly misled by statistical tables. He would have been told that only 4.2% of law students were women and that, looking backwards, the numbers had barely budged over the preceding decades.[190] Fast-forward to the present, and we see that almost half of all law students are women.[191] This monumental change, due largely to the women's movement and anti-discrimination laws, would not have been incorporated into statistical tables.

Great social change continues into the 21st century (does anyone really need to say "President Obama"?). Even today's mortality rates are different from just a few years ago. According to the Centers for Disease Control, "[d]ifferences in mortality between men and women continued to narrow."[192] Similarly, since 1989, the age-adjusted death rates "for the black and white populations have tended toward convergence."[193]

For race- and gender-based statistical tables to accurately foretell the future, therefore, the circumstances which caused the statistical differences would have to continue. The only way to mathematically justify the use of race- and gender-based statistical tables for predicting the future is to assume either that existing discrimination (and its effects) will continue or that the race- and gender-based distinctions are innate and inevitable.[194] Not only are these propositions offensive, pessimistic, and wrong,[195] but also

190. According to the American Bar Association, for the academic year 1963–1964, there were 20,776 first year law students: 19,899 men and 877 women. American Bar Association, Enrollment and Degrees Awarded, 1963–2008, http://www.abanet.org/legaled/statistics/charts/stats%20-%201.pdf (last visited Feb. 6, 2010).

191. For the academic year 2007–2008, 47.3% of the first-year law school class were women; out of 49,082 first year students, 25,864 were men and 23,218 were female. *Id.*

192. Hsiang-Ching Kung et al., *Deaths: Final Data for 2005*, 56 NAT'L VITAL STAT. REP. 1, 2 (2008), *available at* http://www.cdc.gov/nchs/data/nvsr/nvsr56/nvsr56_10.pdf. The age-adjusted death rate for men in 2005 was 40.4% greater than that for women, which was down from being 40.7% greater in 2004. *Id.*

193. *Id.* at 4. According to the CDC, "[d]eath rates declined by 10.6 percent for the black population and by 7.0 percent for the white population between 1989 and 1997, and they have declined by 10.8 percent for the black population and by 8.2 percent for the white population since 1997." *Id.*

194. *See e.g.*, Chamallas, *Civil Rights in Ordinary Tort Cases*, *supra* note 177, at 1455 ("Relying on race and sex-based statistics reinforces the view that race and sex differences are inevitable and enduring, rather than a product of political and social arrangements that are subject to change.").

195. *See, e.g.*, Chamallas, *Questioning the Use*, *supra* note 142, at 75 ("The use of race-based and gender-based tables assumes that the current gender and racial pay gap will continue in the future,

evaluating their likelihood is not within a statistician's skill set. Determining whether discrimination and the effects of past discrimination will be negated by both the legal system and social changes is, most emphatically, not the province of statisticians. The justice system cannot allow itself to be so intimidated by a statistical statement that it overlooks the need to make its own evaluation; indeed, "any decision to use a group-based projection into the future . . . involves normative judgments about the relevant frame of reference and the rate of future change."[196]

For example, as previously discussed, America's fascination with racial analysis often masks, rather than reveals, the truth.[197] While African-Americans, on the average, have a shorter life expectancy than their white counterparts, a large proportion of that difference is due to socioeconomic, not racial, differences. One demographic study found that, "[w]hite residents of urban poor areas have mortality profiles comparable to those of black residents of poor rural areas and blacks nationwide . . . ."[198] The life expectancy of these whites was found to be, in fact, lower than that for "residents of relatively advantaged black urban areas."[199] The socioeconomic factor is disregarded in the life expectancy tables. As Judge Jack Weinstein noted, "[g]ross statistical tables do not answer the question: how does the life expectancy of well-off or middle-class 'African-Americans' compare to that of poor 'African-Americans?'"[200] Thus, he concluded, courts should reject the use of racially based tables that tend to "enforce the negative impacts of lower socio-economic status while ignoring the diversity within populations."[201]

Courts should also recognize that to the extent the differences reflected in the race- and gender-based tables are caused by ongoing discrimination, using those tables reinforces the harm caused by wrongful discrimination. As one court ruled, "it would be inappropriate to incorporate current discrimination resulting in wage differences between the sexes or races or the potential for any future such discrimination into a calculation for

---

despite ongoing legal and institutional efforts to make the workplace more diverse and less discriminatory.").

196. Jennifer B. Wriggins, *Damages in Tort Litigation: Thoughts on Race and Remedies, 1865–2007*, 27 REV. LITIG. 37, 56 (2007).

197. *See supra* notes 115–169 and accompanying text.

198. Geronimus, *supra* note 171, at 234–35; *see also* Joseph J. Sudano & David W. Baker, *Explaining US Racial/Ethnic Disparities in Health Declines and Mortality in Late Middle Age: The Roles of Socioeconomic Status, Health Behaviors, and Health Insurance*, 62 SOC. SCI. & MED. 909, 918 (2006) ("Our results are also consistent with previous studies that have found large 'direct' (or residual) effects of [socioeconomic status] on health that were not explained by differences in health behaviors.").

199. Geronimus, *supra* note 171, at 234–35.

200. McMillan v. City of New York, 253 F.R.D. 247, 252 (E.D.N.Y. 2008).

201. *Id.* at 253.

damages resulting from lost wages."[202] This is especially true in tort cases, where victims have been deprived of their "chance to excel in life beyond predicted statistical averages."[203]

The use of these statistical averages causes harm in a great many ways. The most obvious way is that some injured tort victims receive far less of a remedy than other equally injured tort victims, based solely on their race or gender.[204] Because of a long history of discriminatory treatment, the "explicit use of race-based and sex-based economic data dramatically reduces some damage awards for women and for African-American and Hispanic men."[205] In an infamous 1905 case, the court was faced with determining damages from wrongful death claims for eight claimants whom it described as "white" or "colored."[206] They could not all be treated equally, the court decreed, due to the "difference in the vitality of the two races."[207] Accordingly, the judge, "lowered the awards for the deaths of blacks ten percent more than the awards for the deaths of whites and . . . slashed three of the awards for blacks by forty percent or more."[208]

Racial and gender differences in income continue today. According to a 2007 report by the U.S. Department of Commerce, non-Hispanic white men had annual median earnings of $47,814, while African-American men's annual median earnings were more than twenty-five percent lower, at $34,480.[209] Similarly, the median earnings of women, $32,649, is 77.3% of men's $42,210.[210] For some demographics, the differences are even starker; the average salary for a male Native American is just fifty-eight percent that for white males.[211] Thus, the use of race- and gender-based statistics will

202. Wheeler Tarpeh-Doe v. United States, 771 F. Supp. 427, 455 (D.D.C. 1991), *rev'd sub nom.* Tarpeh-Doe v. United States, 28 F.333d 120 (D.C. Cir. 1994); *see also* United States v. Bedonie, 317 F. Supp. 2d 1285, 1319 (D. Utah 2004), (stating that "[a]s a matter of fairness, the court should exercise its discretion in favor of victims of violent crime and against the possible perpetuation of inappropriate stereotypes"), *aff'd sub nom.* United States v. Serawop, 505 F.3d 1112 (10th Cir. 2007).

203. *Bedonie*, 317 F. Supp. 2d at 1319.

204. Lamb, *supra* note 177, at 302.

205. Chamallas, *Questioning the Use*, *supra* note 142, at 75.

206. The Saginaw & The Hamilton, 139 F. 906, 910 (S.D.N.Y. 1905). The claims resulted from the collision of two steamships, which caused the deaths of both passengers and crewmembers. *Id.*

207. *Id.* at 914.

208. Wriggins, *supra* note 196, at 56.

209. BRUCE H. WEBSTER, JR. & ALEMAYEHU BISHAW, U.S. CENSUS BUREAU, INCOME, EARNINGS, AND POVERTY DATA FROM THE 2006 AMERICAN COMMUNITY SURVEY 15 (2007).

210. *Id.* at 13.

211. United States v. Bedonie, 317 F. Supp. 2d 1285, 1313 (D. Utah 2004), *aff'd sub nom.* United States v. Serawop, 505 F.3d 1112 (10th Cir. 2007).

have the undesirable effect of "reinforcing the underlying social inequalities of our society rather than describing a significant biological difference."[212]

There is an additional social cost, beyond the "perpetuation of inappropriate stereotypes."[213] Assuming that "the deterrent effect of a legal action depends on its ability to raise the cost of the undesirable behavior to the defendant,"[214] it follows that when damages for injuring members of minority groups are lowered, the legal regimen will have the perverse result of encouraging torts against them. Thus, "because it is cheaper to injure poor minority children, there is less incentive for defendants to take measures to clean up toxic hazards in the neighborhoods most affected by lead paint."[215]

A further harm caused by the use of race- and sex-based statistics is analogous to the harm discussed with raced-based DNA testimony: such use places unnecessary emphasis on factors that are both largely irrelevant and have a historical record of justifying irrational discrimination.[216] As one commentator noted, "organizing the statistics around race propels race to the forefront of predictions about individual achievement."[217] The Supreme Court has made a similar observation about the use of gender-based statistics. Because each "individual's life expectancy is based on a number of factors, of which sex is only one[,] . . . [o]ne cannot 'say that an actuarial distinction based entirely on sex is based on any other factor than sex. Sex is exactly what it is based on.'"[218]

There have been significant instances where the fundamental interest in equality has overwhelmed the power of the statistical average. Most notably, perhaps, was the distribution of money from the September 11th Victim Compensation Fund. This fund was established by federal law to provide compensation for those injured or killed as a result of the 9/11

---

212. McMillan v. City of New York, 253 F.R.D. 247, 250 (E.D.N.Y. 2008); *see also* Chamallas, *Civil Rights in Ordinary Tort Cases*, *supra* note 177, at 1439 (stating that reliance on race and gender statistics "saddles nonconforming women and racial minorities with generalizations about their group, the very kind of stereotyping that anti-discrimination laws were meant to prohibit"); Lamb, *supra* note 177, at 304 (stating that the practice of issuing gender-based awards "magnifies the impact of employment discrimination and devalues the earning capacity of injured women, resulting in widely varying damage awards of equally situated men and women for the same injury") (citation omitted).

213. United States v. Serawop, 505 F.3d 1112, 1116 (10th Cir. 2007), quoting lower court decision, *sub nom. Bedonie*, 317 F. Supp. 2d at 1319.

214. Olga N. Sirodoeva-Paxson, *Judicial Removal of Directors: Denial of Directors' License to Steal or Shareholders' Freedom to Vote?*, 50 HASTINGS L.J. 97, 137 (1998).

215. Chamallas, *Civil Rights in Ordinary Tort Cases*, *supra* note 177, at 1441.

216. *See supra* text accompanying notes 150-52.

217. Laura Greenberg, Comment, *Compensating the Lead Poisoned Child: Proposals for Mitigating Discriminatory Damage Awards*, 28 B.C. ENVTL. AFF. L. REV. 429, 447 (2001).

218. Ariz. Governing Comm. for Tax Deferred Annuity & Deferred Comp. Plans v. Norris, 463 U.S. 1073, 1081 (1983) (quoting L.A. Dep't of Water & Power v. Manhart, 435 U.S. 702, 712–13 (1978) (internal quotation marks omitted)).

attacks.[219] A Special Master, Kenneth Feinberg, was appointed to distribute the funds. One of the thornier issues he had to resolve was how to calculate the lost earnings of the victims. In calculating the expected work life for the claimants, the Special Master chose not to "discriminate against women" and elected to utilize the same worklife table for both men and women.[220] The appropriateness of choosing a gender-neutral approach was brought into sharp focus by the overarching purpose of the compensation fund: "[T]o serve as a national expression of unity in the face of a tragedy unique in American history, as well as to help survivors."[221] Thus, the transcendent values of equality and respect for individuals were found to outweigh the persuasive power of statistics in the extraordinary context of compensating for the horrors of 9/11. Those values, though, should also be sufficient in ordinary cases to rebut the need for race- and gender-based statistics.

The Supreme Court took a tentative step toward this goal when it ruled that Title VII of the Civil Rights Act of 1964 prohibits employers from utilizing gender-based statistics in their retirement plans.[222] According to the Court, employers can neither require women to make larger contributions in order to obtain the same monthly pension benefits as men nor offer their employees the option of receiving retirement benefits only with companies that pay lower monthly benefits to a woman than to a man who has made the same contributions.[223] It is irrelevant, the Court explained, whether the sex-segregated actuarial tables actually, "reflect an accurate prediction of the longevity of women as a class."[224] Indeed, "'[e]ven a true generalization about [a] class cannot justify class-based treatment."[225]

---

219. *See* September 11th Victim Compensation Fund of 2001, 28 C.F.R. § 104.1 (2002). The statutory authorization for the fund was contained in Title IV of Public Law 107–42. *Id.*

220. September 11th Victim Compensation Fund, Final Rule, 67 Fed. Reg. 11233, 11238 (Mar. 13, 2002) (codified at 28 C.F.R. § 104). Feinberg chose to apply the worklife table for "males" toall claimants. *Id.*

221. Michael I. Meyerson, *Losses of Equal Value*, N.Y. TIMES, Mar. 24, 2002, at 4 (week in review); *see also* September 11th Victim Compensation Fund of 2001, Interim Final Rule, 66 Fed. Reg. 66274, 66274 (Dec. 21, 2001) (codified at 28 C.F.R. § 104) ("The September 11th Victim Compensation Fund of 2001 is an unprecedented expression of compassion on the part of the American people to the victims and their families devastated by the horror and tragedy of September 11.").

222. *Norris*, 463 U.S. at 1086. Title VII makes it an unlawful employment practice "to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex or national origin . . . ." 42 U.S.C. § 2000e-2(a)(1) (2006). The Court in *Norris* also said, in dictum, that the use of race-based actuarial tables would be similarly illegal. *Norris*, 463 U.S. at 1086.

223. *Norris*, 463 U.S. at 1086; *Manhart*, 435 U.S. at 711.

224. *Norris*, 463 U.S. at 1084 (quoting *Manhart*, 435 U.S. at 708).

225. *Id.* at 1084 (quoting *Manhart*, 435 U.S. at 708). Thus, the Court said, "the greater costs of

Though these rulings were based on a particular federal statute, the reasoning can be applied equally to a constitutional analysis of the use of race- and gender-based statistics. The Court recognized that the use of race and sex to predict longevity was "flatly inconsistent" with the principles of Title VII, which require "employers to treat their employees as *individuals*, not 'as simply components of a racial, religious, sexual, or national class.'"[226]

Such use of race- and sex-based statistics is equally inconsistent with constitutional norms because the same principle applies: "At the heart of the Constitution's guarantee of equal protection lies the simple command that the Government must treat citizens as individuals, not as simply components of a racial, religious, sexual or national class."[227]

It is for the courts, and not the statisticians, to ensure that this command is obeyed. Courts should require the use of "blended tables," which do not distinguish based on race or gender, when calculating tort damages.[228] Just as insurance companies have elected to stop using race-based statistics due to the "social unacceptability" of such discrimination,[229] so should they cease using gender-based statistics, by force of law, if not voluntarily.[230] If society is ready to transcend the history of race and gender discrimination, we must not permit bigoted numbers to slow our progress.

V. RECLAIMING JUDICIAL RESPONSIBILITY FOR ALLOCATING THE RISK OF ERROR

Mistakes happen. There is no "truth machine" that will tell us with unwavering accuracy the proper result of a medical test, economic prediction, or trial.[231] Ideally, we want to reduce the frequency and degree

---

providing retirement benefits for female employees does not justify the use of a sex-based retirement plan." *Id.* at 1085 n.14.

226. *Id.* at 1083 (quoting *Manhart*, 435 U.S. at 708).

227. Miller v. Johnson, 515 U.S. 900, 911 (1995) (quoting Metro Broad., Inc. v. FCC, 497 U.S. 547, 602 (1990) (O'Connor, J., dissenting) (internal quotation marks omitted); *see also* Gratz v. Bollinger, 539 U.S. 244, 270 (2003) (stating that "[r]acial classifications are simply too pernicious to permit any but the most exact connection between justification and classification" (quoting Fullilove v. Klutznick, 448 U.S. 448, 537 (1980) (Stevens, J., dissenting))); United States v. Virginia, 518 U.S. 515, 531 (1996) ("Parties who seek to defend gender-based government action must demonstrate an 'exceedingly persuasive justification' for that action.").

228. Chamallas, *Civil Rights in Ordinary Tort Cases*, *supra* note 177, at 1468.

229. BERTRAM HARNETT & IRVING I. LESNICK, THE LAW OF LIFE AND HEALTH INSURANCE § 13.03 (2008). *See generally Fair Insurance Practices Act: Hearing Before the S. Comm. on Commerce, Science, and Transportation*, 97th Cong., 2nd Sess. 24 (1982).

230. *See, e.g.*, Hartford Accident & Indem. Co. v. Ins. Comm'r of Pa., 482 A.2d 542, 549 (Pa. 1984) (terming differential treatment between men and women by an insurance company "unfair discrimination").

231. *See, e.g.*, Seth F. Kreimer, *Truth Machines and Consequences: The Light and Dark Sides of "Accuracy" in Criminal Justice*, 60 N.Y.U. ANN. SURV. AM. L. 655, 656–67 (2005).

of inaccurate results, but imperfection is an inescapable result of the human condition. Statisticians, who deal in the art of the probable, have devised a useful way to think about and deal with this inevitability of error.

Suppose, for example, that there was a medical test for determining whether patients had a particular disease, and, in general, a higher test score correlated to an increased likelihood of having the disease. Assume that patients have a range of scores on this test, and a cut-off point is needed for purposes of diagnosis.

There are two situations where the test could be wrong. First, with a "false positive," healthy patients are diagnosed with the disease. Alternatively, with a "false negative," diseased patients are mistakenly termed healthy. In statistics, these would be termed "Type I" and "Type II" errors respectively.[232] No matter which cut-off score you choose, you will make some errors; there is no perfect point for us to choose.[233]

Accordingly, the cut-off is chosen based on a determination as to which kind of error is worse than the other. Raising the cut-off point will result in more false negatives (more afflicted patients declared healthy) but fewer false positives (fewer healthy patients deemed afflicted). Lowering the cut-off point has the opposite effect, causing fewer false negatives (fewer afflicted patients declared healthy), but more false positives (more healthy patients deemed afflicted).

Because both kinds of errors will always occur,[234] the cut-off point chosen for determining the presence of the disease will reflect a value judgment as to which error has more serious consequences. We might prefer to have fewer false positives, a smaller Type I error, for an employment drug test, so that we reduce the number of employees wrongfully accused. For diseases with grave consequences that could be averted only by immediate action (as when a change in diet could avoid retardation during fetal development), we might desire fewer false negatives, a smaller Type II rate, to minimize the possibility that someone with the disease goes undiagnosed.

The choice of the legal standard of proof reflects a similar calculus. As with the inevitably imperfect diagnostic medical test, there is always the possibility that the verdict in a trial will not square with the true facts.

In the criminal context, if we convict someone who is innocent, we have made a Type I error. If we acquit a guilty person, we have made a Type II

232. *See, e.g.*, MICHAEL O. FINKELSTEIN & BRUCE LEVIN, STATISTICS FOR LAWYERS 124–26 (2d ed. 2001).

233. *See, e.g., id.* at 120–22.

234. *Id.* The only time you would not have both types of error would be if 100% of those tested fail or 100% pass.

error. Similarly, in the civil context, finding for the plaintiff where, were the truth fully known, the defendant should prevail is a Type I error; permitting the culpable defendant to win the case would be a Type II error. And we know that errors will be made.

Adjusting the standard of proof affects the frequency of each type of error. Just as raising the cut-off point results in fewer healthy people being diagnosed as diseased, the higher we make the standard of proof, the fewer the innocent people who will be found guilty. The cost, of course, is that more guilty people will be acquitted.

The Constitution requires the highest standard for criminal cases, proof beyond a reasonable doubt, because of "a fundamental value determination of our society that it is far worse to convict an innocent man than to let a guilty man go free."[235] This principle predates the Constitution, as reflected in Blackstone's admonition that English law recognized that it was preferable for ten guilty persons to escape than for one innocent person to be convicted wrongfully.[236] The Supreme Court has explained that this balance reflects the fact that the accused has a far greater stake in a criminal trial than even the Government: "Where one party has at stake an interest of transcending value—as a criminal defendant his liberty—[the] margin of error is reduced as to him by the process of placing on the other party the burden of . . . persuading the factfinder . . . of his guilt beyond a reasonable doubt."[237]

By contrast, in a civil suit between two parties, where the plaintiff alleges that the defendant is responsible for some monetary loss, the preponderance-of-the-evidence standard is used, signifying that "the cost of a mistaken verdict for plaintiff is neither greater nor less than the cost of a mistaken verdict for defendant . . . ."[238] As former Chief Justice (then-Associate Justice) Rehnquist noted, because the preponderance-of-the-

---

235. *In re* Winship, 397 U.S. 358, 372 (1970) (Harlan, J., concurring); *see also* Apprendi v. New Jersey, 530 U.S. 466, 477 (2000) (quoting United States v. Gaudin, 515 U.S. 506, 510 (1995) (stating that a criminal defendant is entitled to "a jury determination that [he] is guilty of every element of the crime with which he is charged, beyond a reasonable doubt")).

236. 4 WILLIAM BLACKSTONE, COMMENTARIES *358 ("[T]he law holds, that it is better that ten guilty persons escape, than that one innocent suffer.").

237. Speiser v. Randall, 357 U.S. 513, 525–26 (1958).

238. D.H. Kaye, *Is Proof of Statistical Significance Relevant?*, 61 WASH. L. REV. 1333, 1361 (1986) [hereinafter Kaye, *Is Proof of Statistical Significance Relevant?*]; *see also* Neil B. Cohen, *The Gatekeeping Role in Civil Litigation and the Abdication of Legal Values in Favor of Scientific Values*, 33 SETON HALL L. REV. 943, 950 (2003) [hereinafter Cohen, *The Gatekeeping Role*] ("[T]he preponderance of the evidence standard suggests that the civil litigation system ascribes essentially equal costs to inaccurately proclaiming a proposition to be demonstrated and to inaccurately declining to proclaim that the proposition has been demonstrated."); Posner, *supra* note 7, at 1504 ("In the typical civil trial, there is no basis for supposing that Type I errors (false positives, such as convicting an innocent person) on average impose higher costs than Type II errors (false negatives, such as an erroneous acquittal).").

evidence standard "allocates the risk of error more or less evenly," it is used whenever "an incorrect finding of fault would produce consequences as undesirable as the consequences that would be produced by an incorrect finding of *no* fault."[239]

In some cases, the Court's evaluation of the undesirability resulting from incorrect findings has led to the utilization of a "middle level of burden of proof"—clear and convincing evidence.[240] The Court has required the Government to prove its case by clear and convincing evidence in civil cases in which governmental action threatened a significant deprivation of liberty, such as civil commitment, deportation, and denaturalization.[241]

This increased burden was selected to ensure that more of the risk of an erroneous decision would be imposed on the Government: "The individual should not be asked to share equally with society the risk of error when the possible injury to the individual is significantly greater than any possible harm to the state."[242]

The question of when an injury is "significantly greater" is a matter for judicial determination. For example, in *Santosky v. Kramer*,[243] the Court struggled with the wrenching issuing of adjudicating the loss of parental rights. In that case, three children were removed from their parents' custody after the local Department of Social Services found evidence of abuse, malnutrition and neglect.[244] The issue for the Supreme Court was to determine the standard of proof the Government needed to establish before permanently terminating parental rights.[245]

A majority of the Justices, citing the "commanding" importance of a parent's interest in raising his or her child, required the use of the "clear and convincing" standard.[246] The preponderance-of-the-evidence standard

239. Santosky v. Kramer, 455 U.S. 745, 788 n.13 (1982) (Rehnquist, J., dissenting). Then-Associate Justice Rehnquist referred to such a situation as occurring "when the social disutility of error *in either direction* is roughly equal . . . ." *Id.*

240. Addington v. Texas, 441 U.S. 418, 431 (1979); *see also Santosky*, 455 U.S. at 756 (terming clear and convincing evidence an "intermediate standard of proof").

241. *See, e.g.*, *Santosky*, 455 U.S. at 769; *Addington*, 441 U.S. at 424–26; Woodby v. INS, 385 U.S. 276, 285 (1966); Chaunt v. United States, 364 U.S. 350, 353 (1960); Schneiderman v. United States, 320 U.S. 118, 125 (1943).

242. *Addington*, 441 U.S. at 427; *see also Santosky*, 455 U.S. 745.

243. *Santosky*, 455 U.S. at 745.

244. *Id.* at 751.

245. *Id.* at 747–48.

246. *Id.* at 758. The Court stated that it was "'plain beyond the need for multiple citation' that a natural parent's 'desire for and right to the companionship, care, custody, and management of his or her children' is an interest far more precious than any property right." *Id.* at 758–59 (quoting Lassiter v. Dep't of Soc. Servs., 452 U.S. 18, 27 (1981) (internal quotation marks omitted)).

would "reflect[] the judgment that society is nearly neutral between erroneous termination of parental rights and erroneous failure to terminate those rights."[247] The Court stated that the preponderance-of-the-evidence "standard that allocates the risk of error nearly equally between those two outcomes does not reflect properly their relative severity."[248]

In dissent, then-Associate Justice Rehnquist argued that the majority had seriously undervalued the harm that would result from erroneously maintaining parental rights: "If the Family Court makes an incorrect factual determination resulting in a failure to terminate a parent-child relationship which rightfully should be ended, the child involved must return either to an abusive home or to the often unstable world of foster care."[249] Therefore, he stated, the two types of errors should be viewed as having equal seriousness, and a "preponderance of the evidence" standard should have been utilized to determine what was best for the children.[250]

A second case, *Cruzan v. Director, Missouri Department of Health*,[251] presented a similar need to allocate the risk of error. *Cruzan* involved parents who wanted to terminate the life support system of their comatose daughter.[252] The Supreme Court permitted the State of Missouri to overrule the desires of the parents unless the latter could prove "by clear and convincing evidence" that their daughter would have wanted to avoid further medical treatment.[253] The Court emphasized the State's great interest in the "protection and preservation of human life," and concluded that: "An erroneous decision to withdraw life-sustaining treatment . . . is not susceptible of correction," but "[a]n erroneous decision not to terminate [could be corrected by] the possibility of subsequent developments such as advancements in medical science, [or] the discovery of new evidence regarding the patient's intent."[254]

This time, Justice Brennan dissented, contending that the Court had undervalued the serious harm by incorrectly rejecting the parent's claim: "An erroneous decision not to terminate life support . . . robs a patient of the very qualities protected by the right to avoid unwanted medical treatment. His own degraded existence is perpetuated; his family's suffering is protracted; the memory he leaves behind becomes more and more distorted."[255]

---

247. *Id.* at 765.
248. *Id.* at 766.
249. *Id.* at 789 (Rehnquist, J., dissenting) (footnote omitted).
250. *See id.* at 786–87 (Rehnquist, J., dissenting).
251. 497 U.S. 261 (1990).
252. *Id.* at 265.
253. *Id.*
254. *Id.* at 283.
255. *Id.* at 320 (Brennan, J., dissenting).

In all of these cases, the Supreme Court has struggled to make a careful, nuanced determination as to whether greater harm resulted from one type of error or the other, and then selected the legal standard that incorporates that determination. What matters for this discussion is not whether one agrees with their ultimate determination in any of these cases. Rather, the point is that it is the courts' task to make the normative determination of the harm that would be caused by both types of erroneous decisions, and to adjust the legal standard accordingly.

Unfortunately, as soon as numbers appear in a case, courts appear to abdicate their policy-making responsibilities. For example, in *Castaneda v. Partida*,[256] the Supreme Court was asked to determine whether the Texas system for selecting members of a grand jury discriminated against Mexican-Americans. The most important evidence for the criminal defendant challenging the system was a comparison of the percentage of Mexican-Americans in Hidalgo County with the percentage on grand juries. After a brief statistical analysis, the Court concluded that any claim that "the jury drawing was random would be suspect to a social scientist."[257] That is, however, the wrong question. Whether the statistics would convince a "social scientist" is largely irrelevant; instead, the Court should be asking whether the statistics are legally relevant to resolving the issue at hand.[258]

That decision must incorporate the consideration of how the justice system should balance the risk of erroneously finding discrimination in a fair system with the risk of erroneously finding neutrality in a discriminatory grand jury selection process. While mathematical analysis is necessary to explain to the court the meaning of the numbers, the value-laden evaluation of the comparative seriousness of the harms involved should never be a mathematical decision.

To understand how courts have forfeited their proper role and how they can reclaim it, we must understand the statistical process of hypothesis testing. For that exploration, I will present another story, called *Heads You Win*,[259] to demonstrate the value judgments implicit in hypothesis testing,

256. 430 U.S. 482 (1977).

257. *Id.* at 496 n.17.

258. Three months after *Castaneda* was decided, the Court, in *Hazelwood School District v. United States*, 433 U.S. 299, 309 n.14 (1977), cited *Castaneda* for the proposition that a finding of two or three standard deviations meant that then the hypothesis that teachers in this case were hired without regard to race "would be suspect." However, the *Hazelwood* Court did not say to whom they would be "suspect." *Id.*

259. For an explanation of coin toss probabilities, see *infra* note 280.

and to show how and why judges must reassert their rights (and obligations) to make those judgments.

Heads You Win

> Sally is an elderly art collector who wants to give away the most valuable piece in her collection to one of her two children. She invites them to her house and tells them that she wants to play a game to decide who gets the painting. She opens a fresh roll of quarters and gives one to her son Charles and one to her daughter Lisa. "The game is simple," she says. "Each of you will flip your quarter ten times. Whoever flips heads the most gets the picture. But, if I catch you cheating, not only will you not get the painting, I will leave you out of my will." Charles and Lisa take turns flipping their coins while Sally marks the number of heads that each child flips. When the final scores are tallied, Charles has won with ten heads, while Lisa only has five. Lisa turns to her mother and shouts, "It's not fair. He can't be that lucky. He must have cheated."

Based on the evidence, how is Sally to proceed? More particularly, what do the numbers tell her about the likelihood that Charles cheated?

The case of *Heads You Win* will turn on how we think about unlikely events. While it may seem counterintuitive, every possible outcome of the ten coin tosses is, on one level, unlikely. Because a fair game might suggest there is an equal fifty-fifty likelihood of flipping heads or tails on a single toss, flipping heads on half the tosses, five heads out of ten, would seem to be a likely event. Such thinking, though, ignores the fact that out of ten tosses there are actually eleven possibilities (ranging from zero heads to ten heads). While five is indeed the most likely of the eleven possibilities, one should actually expect to get five heads out of ten tosses fewer than twenty-five percent of the time. While that is the least unlikely event, variation from that ideal would hardly be unexpected.

Similarly, when examining a sample, whether a statistical survey or the scores some people obtained on a job test, one should expect that the sample will not be a perfect representation of the entire population.[260] Similarly, the mere fact that the price of a stock rises the day after information is released does not prove that the stock increase was noteworthy, let alone linked to the information. There is an inevitable volatility in stock prices that might account for the increase.[261]

---

260. Kaye & Freedman, *Reference Guide on Statistics*, *supra* note 10, at 116 ("[A] sample is unlikely to be a perfect microcosm of the population.").

261. *See* Jonathan R. Macey et al., *Lessons from Financial Economics: Materiality, Reliance, and*

Statisticians have a way of calculating the expected volatility in a sample. The mathematical phrase is "standard deviation," which can be thought of as "a measure of spread, dispersion or variability of a group of numbers."[262] The standard deviation is determined based on the amount each element of the sample differs from the average.[263] While there are precise mathematical formulas for calculating standard deviation, there is great benefit to be derived from considering the plain English meaning of the phrase. The phrase "standard deviation" means that, for every statistical sample, some divergence from the center, from the "normal," is to be expected. It is, indeed, "standard," for any specific result to be somewhat different from another.[264]

But the phrase "standard deviation" also implies something else—not all deviations are standard. A degree of variation is to be expected, but some variations are so extreme as to be "non-standard" and surprising. When those surprising variations occur, it makes statisticians consider the possibility that their original expectations might have been mistaken.

To return to the mathematical definition of "standard deviation," the more standard deviations a result is from the expected result (oft times the mean), the less likely one is to see it. The most common mathematical calculation for standard deviation involves what is known as a "normal distribution."[265] The normal distribution can be thought of as the classic "bell curve," a symmetric distribution with the highest total occurring in the

---

Extending the Reach of Basic v. Levinson, 77 VA. L. REV. 1017, 1036 (1991) ("To test for such statistically significant returns, it is necessary to account for the usual volatility of returns, which varies across firms and over time.").

262. DAVID C. BALDUS & JAMES W. L. COLE, STATISTICAL PROOF OF DISCRIMINATION 359 (1980).

263. Kaye & Freedman, *Reference Guide on Statistics*, *supra* note 10, at 174. Depending on the statistical test being utilized, there are numerous formulas for standard deviations. *See* ARTHUR M. GLENBERG, LEARNING FROM DATA: AN INTRODUCTION TO STATISTICAL REASONING 66 (2d ed. 1996).

264. Kaye & Freedman, *Reference Guide on Statistics*, *supra* note 10, at 117 ("An estimate based on a sample is likely to be off the mark, at least by a little, due to random error.").

265. The normal distribution shows the probability of a continuous range of possible occurrences. *See* FINKELSTEIN & LEVIN, *supra* note 232, at 113. It is a perfectly symmetric curve around the mean, with probabilities above the mean identical to corresponding probabilities below the mean. *See id.* Even though life is usually not neat enough to fall into a normal distribution, statisticians are often able to work with data in such a way that it approximates the normal distribution. *See* WAYNE C. CURTIS, STATISTICAL CONCEPTS FOR ATTORNEYS 71 (1983). In fact, "[a]lthough stock returns are actually not distributed normally, researchers have shown that the normal distribution is a good approximation for event study estimations." Macey et al., *supra* note 261, at 1039. Despite the seeming prevalence of statistical use of the normal distribution, other distributions, such as binomial distributions, are also utilized by statisticians. *See, e.g.*, D. G. REES, ESSENTIAL STATISTICS 73–74 (4th ed. 2001).

middle and smaller totals occurring as one goes to either extreme.[266] With a normal distribution, a little more than two-thirds of all results will be within one standard deviation of the mean.[267] Slightly fewer than one-third of all results are more than one standard deviation away.[268] Thus, the mere fact that results occur one standard deviation from the mean is not particularly shocking. An occurrence with a one-third probability, say rolling a one or two with a single die, would not raise any eyebrows.

Once we move more than two standard deviations from the mean, however, suspicions often rise. Because the probability is approximately 95.5% that a randomly selected result will fall within two standard deviations of the mean, there is a less than five percent chance of seeing a result that is more than two standard deviations from the mean.[269]

In *Heads You Win*, the task for Sally is to determine whether the specific result Charles obtained, ten heads out of ten tosses, is so surprising as to raise the specter that the tosses were not fair. To make that determination, statisticians might suggest that she use hypothesis testing to quantify this intuition.

Perhaps the most remarkable aspect of hypothesis testing is that usually it does not actually test the proposition in which one is most interested.[270] Rather, hypothesis testing generally examines the likelihood that the opposite of what you care about is true.[271] For example, the question Sally needs to decide is whether Charles cheated. No mathematical tool exists for determining this directly. Instead, an examination of the data through hypothesis testing is a journey in indirection. The focus of the hypothesis test would be an evaluation of the probability of seeing Charles's result (ten heads) had he not cheated.[272] Specifically, all that a hypothesis test can show is "whether an observed result is so unlikely to have occurred by chance alone that it is reasonable to attribute the result to something else."[273] We are left with a double negative. The test cannot tell us if Charles probably cheated; at best, it can tell us something like "Charles probably did not 'not cheat.'"[274]

---

266. *See* FINKELSTEIN & LEVIN, *supra* note 232, at 113.

267. More precisely, the probability that a randomly selected value will be within one standard deviation of the mean is about 68.3%. *See* CURTIS, *supra* note 265, at 73.

268. The probability that a randomly selected value will be greater than one standard deviation of the mean is about 31.7%. *Id.*

269. *See id.* Results more than three standard deviations from the mean are even rarer. The probability of being more than three standard deviations from the mean in a normal distribution is less than one percent, approximately .3%. *Id.*

270. *See* CURTIS, *supra* note 265, at 119.

271. *Id.*

272. Note that this is not the same as an evaluation of the probability of there being no cheating.

273. Kaye, *Is Proof of Statistical Significance Relevant?*, *supra* note 238, at 1333.

274. *See* Kaye & Freedman, *Reference Guide on Statistics*, *supra* note 10, at 122 ("Regrettably,

The first step in hypothesis testing is to identify what we want to ascertain, namely the proposition that Charles's result is a product of cheating. This is sometimes described, ironically, as the "alternate hypothesis."[275] If there were multiple possible causes for the result, i.e. maybe the coin was not fair to begin with or maybe Sally herself doctored the coin, there would be several ways to phrase the alternate hypothesis.[276] For this example, assume that if the process was not fair, Charles cheated.

Sally would then label the opposite proposition, that Charles's result of ten heads is a product of random chance, as the "null hypothesis."[277] Generally, the null hypothesis is a statement that "differences in the sample are due to the luck of the draw."[278] In logical terms, we can say that if the null hypothesis is false, the alternate hypothesis is accurate. Thus, the goal of hypothesis testing is to attempt to disprove the null hypothesis.

To do that, Sally would calculate what is known as a "P-value." For different types of data, different statistical tests would be used to determine the P-value, but all are generally designed to answer one question: What is the probability of seeing a result as "extreme" as the result actually seen if the null hypothesis were true?[279] Sally would discover that the P-value for Charles's result is less than one percent, only about 0.00098.[280]

---

multiple negatives are involved here. A statistical test is essentially an argument by contradiction.").

275. The "alternate hypothesis" is also termed H_1. REES, *supra* note 265, at 141.

276. *See, e.g.*, Kaye, *Is Proof of Statistical Significance Relevant?*, *supra* note 238, at 1355 ("[T]here are always other alternatives besides the one the statistician identifies as H_1 in formulating the test."). For a discussion of the importance of correctly describing the alternate hypothesis, see Daniel L. Rubinfeld, *Econometrics in the Courtroom*, 85 COLUM. L. REV. 1048, 1055 (1985) ("[T]he form of the alternative hypothesis can affect the conclusion that one reaches from the statistical analysis.").

277. The "null hypothesis" is also termed H_0. REES, *supra* note 265, at 141.

278. Kaye & Freedman, *Reference Guide on Statistics*, *supra* note 10, at 122. Another way of describing the null hypothesis is to state that "the difference observed in the data is then just due to sampling error." *Id.* at 173.

279. Kaye & Freedman, *Reference Guide on Statistics*, *supra* note 10, at 117.

280. *See* Kaye, *supra* note 238, at 1350. If you toss a coin, there are only two possible outcomes for each toss. If you toss it ten times, the total number of possible outcomes is 2^10 = 1024. *Id.* Of these, there are 252 ways to get exactly five heads (only the first five tosses could be heads, every other toss could be heads, etc . . .). Thus, the probability of tossing exactly ten heads is 1/1024 = 0.00098. *Id.*

The following chart shows the probability of tossing each quantity of heads:

0 heads: 1/1024 = 0.00098
1 head: 10/1024 = 0.00977
2 heads: 45/1024 = 0.04395
3 heads: 120/1024 = 0.11719
4 heads: 210/1024 = 0.20508
5 heads: 252/1024 = 0.24609

What does that very small number tell us? It means that it would be extremely unlikely to see as many as ten heads out of ten tosses as a random result of a fair process. A statistician would find this result "statistically significant," and therefore "reject" the null hypothesis that Charles's ten heads were the result of pure chance. Because the probability of seeing ten heads for ten tosses is so small under the null hypothesis, Charles's result would "be strong evidence" that the coin toss was not fair.[281]

To understand what it means to say that the probability was "so small" that it would be strong evidence of an unfair toss, we can change the plot of *Heads You Win.* Suppose that instead of ten tosses, Sally asked for one hundred tosses and Charles had tossed sixty heads out of his one hundred tosses. If his sister again accused him of cheating, the analysis would need be slightly different. We would not use a null hypothesis based on exactly sixty heads appearing as a product of random chance. If the allegation is that "he can't be that lucky," we need to consider the probability of a random person being precisely *that* lucky (sixty heads out of one hundred flips) or even luckier (tossing sixty-one heads out of one hundred flips, tossing ninety heads, etc.). In fact, to see how "extreme" sixty heads is, statisticians would want to calculate how likely it would be to obtain sixty or more of either heads or tails. Thus, the new null hypothesis is that tossing sixty or more heads or sixty or more tails out of one hundred tosses is a product of random chance. In calculating this new P-value, Sally would learn that the probability of seeing a result as "extreme" as sixty heads due to random chance is about 5.69%.[282]

The key question is whether this result, the probability of seeing a result as extreme as sixty heads out of one hundred tosses if the null hypothesis were true, is so small that we should suspect that the coin toss was not fair. According to statisticians, the magic number for a P-value is .05 (i.e. a five percent probability), which is approximately two standard deviations. If the P-value is greater than .05, the results are deemed to be not "statistically significant," and hence not sufficient to disprove the null hypothesis.[283] A

---

6 heads: 210/1024 = 0.20508
7 heads: 120/1024 = 0.11719
8 heads: 45/1024 = 0.04395
9 heads: 10/1024 = 0.00977
10 heads: 1/1024 = 0.00098

281. Kaye, *Is Proof of Statistical Significance Relevant?*, *supra* note 238, at 1350.

282. The difference between "extreme" meaning sixty or more heads, and "extreme" meaning sixty or more heads *or* sixty or more tails, is the difference between the one-tail test and the two-tail test. *See infra* text accompanying notes 415–26.

283. "According to the current paradigm, an observation is deemed 'statistically significant' (test hypothesis rejected, null hypothesis given consideration) if the p-value is less than 0.05; an observation is deemed 'not significant' (test hypothesis 'not rejected' or accepted) if the p-value is greater than 0.05." David Egilman et al., *Proving Causation: The Use and Abuse of Medical and Scientific Evidence Inside the Courtroom—An Epidemiologist's Critique of the Judicial*

test where the P-value must be less than .05 to be deemed statistically significant is said by statisticians to have a .05 significance level. Another concept, termed either the "confidence level" or "confidence coefficient," is defined as being equal to one minus the significance level; thus, statisticians say that the confidence level for such a test is ninety-five percent.

With such a confidence level, because the P-value is greater than .05, Sally would not "reject" the hypothesis that the Charles's result is from random chance.[284] This would not mean we had proven that Charles's sixty heads were actually the result of random chance; it would merely announce that the data was not inconsistent with the supposition that the result was due to the luck of the draw.[285] Thus, a statistician would conclude that, because we obtained a P-value of greater than .05, we cannot say that that the null hypothesis is probably false, and therefore, cannot say that the alternate hypothesis of cheating is probably true.

The overwhelming majority of courts have accepted as dogma a rule that any P-value greater than either .05 or less than two standard deviations is not sufficient to disprove a null hypothesis of random chance.[286] The Supreme Court, too, has indicated that it leans towards such an approach. In *Castaneda v. Partida*,[287] the Court was attempting to determine whether Mexican-Americans had been underrepresented in Texas grand juries. In comparing the percentage of Mexican-Americans eligible to serve with those who did serve, the Court stated: "As a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the

---

Interpretation of the Daubert *Ruling*, 58 FOOD & DRUG L.J. 223, 240 (2003).

284. *See, e.g.*, Neil B. Cohen, *Confidence in Probability: Burdens of Persuasion in a World of Imperfect Knowledge*, 60 N.Y.U. L. REV. 385 (1985) [hereinafter Cohen, *Confidence in Probability*].

285. Some courts have failed to understand the limited meaning of a failure to find statistically significant results. One court stated, "[i]f a significant difference is found, the null hypothesis is rejected. If a significant difference is not found, the null hypothesis is accepted." Merrell Dow Pharms., Inc. v. Havner, 953 S.W.2d 706, 722 (Tex. 1997). That is incorrect; if a significant difference is not found, the null hypothesis is "not rejected." We cannot say if the null hypothesis true.

286. *See, e.g.*, Marcel C. Garaud, Comment, *Legal Standards and Statistical Proof in Title VII Litigation: In Search of a Coherent Disparate Impact Model*, 139 U. PA. L. REV. 455, 467 (1990) ("In fact, courts have applied a 95% confidence coefficient corresponding to a 5% significance level cut-off in disparate impact cases."); *see also* Arnold Barnett, *An Underestimated Threat to Multiple Regression Analyses Used in Job Discrimination Cases*, 5 INDUS. REL. L.J. 156, 168 (1982) ("The most common rule is that evidence is compelling if and only if the probability the pattern obtained would have arisen by chance alone does not exceed five percent.").

287. 430 U.S. 482 (1977).

jury drawing was random would be suspect to a social scientist."[288] As some have noted, the fact that the actual number of standard deviations in that case was "approximately [twenty-nine] standard deviations,"[289] combined with "the casualness of the Court's language in the footnote," indicates that the Court did not intend to fix a mandatory level for statistical significance.[290]

Nonetheless, the vast majority of courts considering this question have opted for the security of replicating the classical statistical model. Without evaluating whether the concept of statistical significance is equivalent to the concept of legal significance,[291] courts have generally appropriated the traditional statistical world view: "Social scientists, and in turn the courts, have adopted two standard deviations as a threshold measure of statistical significance."[292]

In one case, African-American employees attempted to prove discrimination by showing that whites had been promoted at a much higher

---

288. *Id.* at 497 n.17. Later that year, the Supreme Court reaffirmed this statement. In *Hazelwood Sch. Dist. v. United States*, 433 U.S. 299, 309 n.14 (1977), the Court, in deciding whether a school district had discriminated against African-Americans in its hiring of teachers, discussed the statistical comparison between the racial compositions of the defendant school district's teaching staff with the public school teacher population in the relevant labor market. *Id.* The Court reiterated its earlier comments: "The Court in *Castaneda* noted that '[a]s a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations,' then the hypothesis that teachers were hired without regard to race would be suspect." *Id.* (quoting *Castaneda,* 430 U.S. at 497 n.17).

289. *Castaneda,* 430 U.S. at 497 n.17 (stating that, "[t]he 11-year data here reflect a difference between the expected and observed number of Mexican-Americans of approximately 29 standard deviations.").

290. Segar v. Smith, 738 F.2d 1249, 1283 n.28 (D.C. Cir. 1984).

291. *See generally* Kaye, *Is Proof of Statistical Significance Relevant?*, *supra* note 238.

292. Dobbs-Weinstein v. Vanderbilt Univ., 1 F. Supp. 2d 783, 803 (M.D. Tenn. 1998); *see also* Davis v. New York City Hous. Auth., 60 F. Supp. 2d 220, 239 (S.D.N.Y. 1999), *aff'd in relevant part*, 278 F.3d 64 (2d Cir. 2002) ("Courts have frequently adopted a standard of two to three standard deviations as constituting statistical significance."). *See, e.g.*, Jones v. GPU, Inc., 234 F.R.D. 82, 95 n.53 (E.D. Pa. 2005) ("Standard deviation units measure statistical significance. 1.96 standard deviation units refers to the level of statistical disparity required to demonstrate legal statistical significance using a two-tailed test."); Smith v. Xerox Corp., 196 F.3d 358, 366 (2d Cir. N.Y. 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. Courts generally consider this level of significance sufficient to warrant an inference of discrimination."); Government v. Penn, 838 F. Supp. 1054, 1070 (D.V.I. 1993) (stating that "statistically significant" refers to "at least a 95 percent probability . . . ."); United States v. Lansdowne Swim Club, 713 F. Supp. 785, 809 (E.D. Pa. 1989) ("[s]tandard deviation of greater than two or three excludes chance"); Frazier v. Consolidated Rail Corp., 851 F.2d 1447, 1452 (D.C. Cir. 1988) ("The question—the legal question—is what degree of certainty the courts require for a *prima facie* case to be established. The 5% level . . . is commonly accepted among statisticians as an acceptable degree of uncertainty . . . ."); Palmer v. Shultz, 815 F.2d 84, 96 (D.C. Cir. 1987) ("[S]tatistical evidence must meet the 5% level . . . for it alone to establish a prima facie case under Title VII."); Whelan v. Merrell-Dow Pharms., Inc., 117 F.R.D. 299, 304 (D.D.C. 1987) ("[S]tatistical evidence is admissible only if that evidence is statistically significant at the 95% confidence level . . . .").

rate.[293] Out of twenty-two total promotions, only five went to blacks.[294] Because forty-two percent of the eligible workforce was black, it would be expected, had there been no discrimination, that forty-two percent of twenty-two, or 9.24, of those promoted would have been black.[295] To determine whether the difference between the "expected" 9.24 and the "actual" five was statistically significant, the court used a null hypothesis of "no discrimination," and calculated how likely it would be to see that more whites than blacks were hired if the difference were due entirely to random chance.[296] The court announced that when "the difference is less than 2 standard deviations, it is not statistically significant."[297] Because the difference in this case amounted to 1.84 standard deviations, it did not reach the two standard deviation threshold.[298] "Consequently," the court concluded, "the plaintiffs failed to prove a prima facie case of discrimination."[299]

Obviously, the court's decision to use two standard deviations was the critical element in its ruling on the existence of differential treatment. Yet, an accurate understanding of this benchmark reveals that, once again, mathematically ignorant judges have ceded their responsibility to make normative policy judgments.

As a starting point, it must be noted that the origin of the .05 significance level was intuition, rather than rigorous mathematics. Statisticians, working on researching various industrial and agricultural problems, were attempting to show how their mathematical tools could help

---

293. Anderson v Douglas & Lomason Co., 26 F.3d 1277 (5th Cir. 1994).
294. *Id.* at 1292 n.26.
295. *Id.*
296. *Id.*
297. *Id.*
298. *Id.* The court used the following formula to calculate the number of standard deviations:

$$\text{Number of } S/D = \frac{(O-NP)}{\sqrt{NP(1-NP)}}.$$

S/D = Standard Deviations
O = Actual number of blacks who received a promotion
N = Number of workers who received a promotion
P = Probability of a black being promoted from the relevant population

$$\text{Thus, } \frac{5-(22 \times 42\%)}{2.30} = -1.84.$$

Id.
299. *Id.*

point to solutions that, when repeated, would prove successful.[300] The founder of modern statistics,[301] R.A. Fisher, wrote, almost cavalierly, "it is convenient to draw the line at about the level at which we can say: 'Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials.'"[302] Fisher actually acknowledged that choice was a subjective choice, not mandated by either science or math.[303] Nonetheless, Fisher's choice quickly became the gold standard for statisticians.

Far more troubling than the birth of the .05 standard, is that its use leads to a skewed balancing of the risks of different types of errors. Returning to our modified version of *Heads You Win*, if Sally wrongfully concluded that Charles had cheated when his result was merely the result of random chance, she would be making the Type I error of incorrectly condemning the innocent. If, on the other hand, she were to decide mistakenly that Charles's sixty heads were the result of random chance, though in reality he had cheated, Sally would be making the Type II error of incorrectly exonerating the culpable.[304]

As Fisher wanted, the use of the .05 significance level results in decision-making in which the probability is extremely small that one will erroneously reject the null hypothesis, for example, by believing that the result was caused by an unfair system when it was merely the product of random chance. When people in Sally's position conclude that someone was not culpable, they will be wrong only five percent of the time.

Put slightly differently, if we imagined a large series of ten coin tosses, the hypothesis of fairness would sometimes be rejected and sometimes not rejected. With a P-value of .05, if we rejected the hypothesis of fairness one

---

300. *See* KENNETH J. ROTHMAN ET AL., MODERN EPIDEMIOLOGY 151 (3d ed. 2008) ("The preoccupation with significance testing derives from the research interests of statisticians who pioneered the development of statistical theory in the early 20th century. Their research problems were primarily industrial and agricultural, and they typically involved randomized experiments . . . that formed the basis for a choice between two or more alternative courses of action. Such experiments were designed to produce results that would enable a decision to be made, and the statistical methods employed were intended to facilitate decision making."). *See generally* Egilman et al., *supra* note 283.

301. *See* C. Radhakrishna Rao, *R.A. Fisher: The Founder of Modern Statistics*, 7 STAT. SCI. 34 (1992). Indeed, Fisher is credited with coining the very word "statistic." Leonard J. Savage, *On Rereading R. A. Fisher*, 4 ANNALS STAT., 441, 452 (1976).

302. Ronald A. Fisher, *The Arrangement of Field Experiments*, *in* BREAKTHROUGHS IN STATISTICS: FOUNDATIONS AND BASIC THEORY 83 (Samuel Kotz et al. eds., 1993).

303. After discussing other possible standards of significance, Fisher declared: "Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level." *Id.*

304. A Type I error has also been characterized as a "false inculpation" while a Type II error is termed a "false exculpation." Cohen, *Confidence in Probability*, *supra* note 284, at 410.

hundred times, we would be making a mistake of assuming a fair coin was not fair only five times.

Perhaps surprisingly, the .05 significance level does not tell us about the other sort of error—the Type II error of failing to reject the hypothesis of fairness when in fact the coin toss was not fair. In our large series of coin tosses, we do not know how often people in Sally's position fail to identify a truly unfair coin.

The exclusive focus on reducing Type I errors does more than mask the existence of Type II errors. Even more problematically, the more we strive to reduce Type I errors, the greater will be the risk of Type II errors.[305] With hypothesis testing it is impossible to reduce the risk of both Type I and Type II errors; thus, a decrease in one results in an increase of the other.[306] We can see this intuitively with the stringent "beyond a reasonable doubt" standard; fewer innocent people are convicted (Type I error), but more guilty parties are acquitted (Type II error).[307] In general, Type I and Type II risks are "inversely related," since by reducing one we tend to increase the other.[308]

The precise mathematical relationship between Type I and Type II errors, however, is not simple to calculate.[309] While there is a direct relationship between the two, it is not a linear relationship; while an increase in one type of error will lead to a decrease in the other, it will not necessarily be of an equal amount.[310]

There is a statistical measure of the ability of a test to prevent Type II errors. Statisticians use the word "power" to describe the probability of

---

305. *See* Macey et al., *supra* note 261, at 1041 ("The tradeoff, however, is that while the higher significance level reduces Type 1 errors, it also increases the probability of Type 2 errors.").

306. In some situations, a researcher can reduce both Type I and Type II errors simultaneously by increasing sample size. RON N. FORTHOFER ET AL., BIOSTATISTICS: A GUIDE TO DESIGN, ANALYSIS, AND DISCOVERY 240 (2007). Such an increase, however, may greatly increase the cost of an experiment. GEOFFREY KEPPEL ET AL., INTRODUCTION TO DESIGN AND ANALYSIS 195 (2d ed. 1992). Moreover, it will frequently be impossible to alter the number of events available for evaluation. *Id.*

307. *See* Cohen, *Confidence in Probability*, *supra* note 284, at 411 n.113 ("It is easy to see, for any given quantity of data, that a rule of decision that decreases the likelihood of Type I (false inculpation) errors will increase the likelihood of Type II (false exculpation) errors, and vice versa."). *See generally* Richard A. Posner, *An Economic Approach to Legal Procedure and Judicial Administration*, 2 J. LEGAL STUD. 399, 408–15 (1973).

308. Cohen, *Confidence in Probability*, *supra* note 284, at 411.

309. *See, e.g.*, DeLuca v. Merrell Dow Pharms., Inc., 911 F.2d 941, 947 (3d Cir. 1990) ("Unfortunately, the relationship between type one error and type two error is not simple . . . .").

310. *See, e.g.*, Cohen, *Confidence in Probability*, *supra* note 284, at 411 ("Although the two risks are inversely related in that increasing one decreases the other, they are not simple complements—that is, they do not add up to one.") (citation omitted).

properly rejecting the null hypothesis when the alternative hypothesis is correct.[311] A high power means fewer Type II errors. It turns out that tests using the .05 significance level, while very effective at preventing Type I errors, generally have low power; they are not particularly good at preventing Type II errors.

Professors David Kaye and David Freedman provide a useful example to demonstrate this tradeoff.[312] An employer plans to use an examination to select trainees. To see whether there is a disparate impact, the employer administers the exam to a sample of fifty men and fifty women drawn at random from the pool of job applicants. If the null hypothesis is that men and women pass the test at equal rates, and the P-value is set at .05, courts will mistakenly find a disparate impact—that is, incorrectly reject the null hypothesis—no more than five times out of one hundred. But what if, in reality, the test does have a disparate impact such that fifty-five percent of the men would pass, but only forty-five percent of the women would. In such a case, courts would mistakenly find no disparate impact more than eighty times out of one hundred.[313] Not only would the court be wrong more than half the time, such a statistical analysis would result in a probability of an incorrect exoneration that is more than sixteen times the probability of an incorrect condemnation that statisticians would be willing to accept.

The disparity may actually be even worse than that. Assume that tests and other hiring practices in this particular industry have always favored men, such that we may consider the probability that the test favored women to be negligible. Thus, any disparate impact would find men doing better than women. The P-value of .05 would then result in courts mistakenly finding a disparate impact, that is, incorrectly rejecting the null hypothesis, no more than 2.5 times out of 100.[314] In such a case, the probability of an incorrect exoneration is more than thirty-two times the probability of an incorrect condemnation.

This level of difference between Type I and Type II errors is typical of hypothesis testing in general. One statistical model demonstrated that, for a hypothetical employment discrimination case, when the risk of incorrectly

311. Kaye & Freedman, *Reference Guide on Statistics*, *supra* note 10, at 125 n.144; *see also* Richard Goldstein, *Two Types of Statistical Errors in Employment Discrimination Cases*, 26 JURIMETRICS J. 32, 34 (1985) ("Power is the probability of *not* making a Type 2 error. In other words, power is the probability of correctly rejecting the null hypothesis.").

312. Kaye & Freedman, *Reference Guide on Statistics*, *supra* note 10, at 121–26, 156–59.

313. In the real world, statisticians, when they focus on Type II errors at all, are generally quite willing to risk many more Type II errors than Type I. As one text on biostatistics noted, the "value of 0.20 for [risk of Type II error] is used frequently in the literature." FORTHOFER, *supra* note 306, at 218.

314. The concept that results are only likely to be "extreme" in one direction, i.e. hiring too few women hired but not too few men, is captured by the concept of a "one-tailed test." *See infra* text accompanying notes 415–26.

condemning the innocent employer (Type I error) was set at .05, the risk of incorrectly exonerating the discriminatory employer (Type II error) was approximately fifty percent, a ten-times greater risk.[315]

Such a disparity has severe practical consequences for our justice system. Innocent employers will lose only one time out of twenty, while injured employees lose half of the time. There may well be legitimate policy reasons for so allocating the risk of errors in certain circumstances, but courts rarely engage in this analysis. Far too many judges have been unable to see the policy choices inherent in the numbers. Instead, they meekly accept the .05 significance level as beyond their capacity to alter.[316]

The consequence of this mathematical illiteracy is what has been termed "an arbitrary balancing of the disutilities, or regrets, of Type I and Type II errors."[317] But such a balancing should reflect a comparison of the social harms associated with each type of error. While judges have yet to appreciate the need to make this comparison for the justice system, statisticians realized early on the need to make their own value judgments that reflect the cost of different types of errors in their very dissimilar field.[318]

As one of R.A. Fisher's students would later remark, Fisher "vehemently denied" the importance of Type II errors for the work of statisticians.[319] Fisher himself wrote:

> The notion of an error of the so-called "second kind," due to accepting the null hypothesis "when it is false" may then be given a meaning in reference to the quantity to be estimated. It has no meaning with respect to simple tests of significance, in which the

315. *See* John M. Dawson, *Probabilities and Prejudice in Establishing Statistical Inferences*, 13 JURIMETRICS J. 191, 201–09 (1973); *see also* Goldstein, *supra* 311 (finding in one example that the risk of a Type II error was almost fifty percent (0.4919) for a significance level of ninety-five percent). Traditionally, "a twenty percent risk of a Type II error is deemed acceptable by statisticians . . . ." Michelle M. Mello, *Using Statistical Evidence to Prove the Malpractice Standard of Care: Bridging Legal, Clinical, and Statistical Thinking*, 37 WAKE FOREST L. REV. 821, 841 n.54 (2002). *See generally* Cohen, *Confidence in Probability*, *supra* note 284, at 410–12.

316. As the Texas Supreme Court declared: "We think it unwise to depart from the methodology that is at present generally accepted among epidemiologists . . . . Accordingly, we should not widen the boundaries at which courts will acknowledge a statistically significant association beyond the 95% level . . . ." Merrell Dow Pharms., Inc. v. Havner, 953 S.W.2d 706, 724 (Tex. 1997).

317. Cohen, *Confidence in Probability*, *supra* note 284, at 412 (citation and internal quotation marks omitted).

318. *Id.*

319. Savage, *supra* note 301, at 441–500.

only available expectations are those which flow from the null hypothesis being true.[320]

One of the reasons for the statistics community's general indifference to Type II errors is that, in the world of scientific evaluations for which these tests were designed, such errors are not final. Failing to reject the null hypothesis of "no effect" is not the same as accepting the proposition that a given substance or technique actually had no effect. The statistical failure to reject the null hypothesis is nothing more than "cause to reserve judgment on the proposition."[321] As R.A. Fisher explained: "[I]t should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."[322]

In this light, the differing harms for a statistician or scientist of a Type I versus Type II error become obvious. When a Type I error is made, and a null hypothesis of "no effect" is incorrectly rejected, a scientist has mistakenly declared "a predictive rule of nature rests on these test results."[323] By contrast, when a Type II error is made and an erroneous null hypothesis of "no effect" is not rejected, no definitive statement has been made. Moreover, scientists are free to conduct countless further studies which might reveal the truth.[324] Thus, "the .05 level reflects the social scientist's conservatism with respect to Type I error."[325]

Judges should have a markedly different view of the comparative costs of Type I and Type II errors. A judge, unlike a scientist, is not "just deferring decision until more research becomes available. Rather, a judge is

---

320. RONALD A. FISHER, THE DESIGN OF EXPERIMENTS 17 (8th ed. 1966).

321. Mello, *supra* note 315, at 842.

322. FISHER, *supra* note 320, at 18.

323. Margaret G. Farrell, Daubert v. Merrell Dow Pharmaceuticals, Inc.: *Epistemiology and Legal Process*, 15 CARDOZO L. REV. 2183, 2210 (1994). As R.A. Fisher described this world view: "A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance." RONALD A. FISHER, *The Arrangement of Field Experiments*, *in* BREAKTHROUGHS IN STATISTICS: FOUNDATIONS AND BASIC THEORY 83 (Samuel Kotz et al. eds., 1993) (emphasis in original).

324. As one commentator noted, "the time frame of science is relatively open-ended." Peter H. Schuck, *Multi-Culturalism Redux: Science, Law, and Politics*, 11 YALE L. & POL'Y REV. 1, 17 (1993).

325. Richard Lempert, *Statistics in the Courtroom: Building on Rubinfeld*, 85 COLUM. L. REV. 1098, 1099 (1985); *see also* Lucinda M. Finley, *Guarding the Gate to the Courthouse: How Trial Judges Are Using Their Evidentiary Screening Role to Remake Tort Causation Rules*, 49 DEPAUL L. REV. 335, 364 (1999) ("Indeed, epidemiology is so inherently conservative in its reluctance to abandon the null hypothesis that it is far more willing to tolerate false negatives—the rejection of a causal association when one may actually exist—than false positives—the attribution of an association when one does not exist."); Kaye, *Is Proof of Statistical Significance Relevant?*, *supra* note 238, at 1343 ("[S]ocial scientists adopted the methods and conventions of others who were concerned primarily with problems in biology.").

selecting a specific course of action that definitively resolves important social and legal rights . . . ."[326] Most significantly, a judicial Type II error, failing to reject an erroneous null hypothesis of "no effect" or "no discrimination," is actually a legal acceptance of that false premise.[327] When such a Type II error occurs, a truly-harmed plaintiff is denied relief.

Nonetheless, courts should be somewhat cautious about finding liability every time numbers vary from expected values. When one contemplates all of the events that individually are unlikely and considers them all together, "it would be very unlikely for unlikely events not to occur."[328] Accordingly, unless one would find a result truly surprising, we should not reject the possibility that it was the product of random chance.

Thus, in our modified version of *Heads You Win*, Sally might not be willing to accuse Charles of cheating just because he obtained sixty heads out of one hundred tosses. With a greater than five percent likelihood that his result occurred due to random chance,[329] she might well prefer avoiding the risk of making the Type I error of condemning the innocent and risk the Type II error of exonerating the culpable. One reason for such a calibration is that there was no evidence of Charles's malfeasance except for the numbers themselves. Some have termed the situation where the only evidence is statistical as "naked statistical evidence."[330]

The most well-known hypothetical involving naked statistical evidence is the case of the "Blue Bus," in which a driver is struck by a bus, and the only evidence available is that eighty percent of the buses that run on the road where the accident occurred are operated by the Blue Bus Company.[331]

---

326. Finley, *supra* note 325, at 366.

327. *See, e.g.*, Mello, *supra* note 315, at 842 ("A legal adjudicator, in contrast, will treat that finding as effectively establishing the null hypothesis.").

328. JOHN ALLEN PAULOS, INNUMERACY: MATHEMATICAL ILLITERACY AND ITS CONSEQUENCES 28 (1988); *see also* Savage, *supra* note 301, at 473 ("The logic of 'something unusual' is very puzzling, because of course in almost any experiment, whatever happens will have astronomically small probability under any hypothesis. If, for example, we flipped a coin 100 times to investigate whether the coin is fair, all sequences have the extremely small probability of 2^{-100} if the coin is fair, so something unusual is bound to happen.").

329. *See supra* text accompanying note 282.

330. *See* David H. Kaye, *Naked Statistical Evidence*, 89 YALE L.J. 601 (1980) (book review); *see also* Richard Lempert, *The New Evidence of Scholarship: Analyzing the Process of Proof*, 66 B.U. L. REV. 439, 460 (1986).

331. Charles Nesson, *The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts*, 98 HARV. L. REV. 1357, 1378 (1985). A related hypothetical is the "Gatecrasher" case, in which the owner of a rodeo discovers that 501 spectators did not pay to enter, while 499 did, and wants to sue all 1000 spectators on the grounds that it is, statistically speaking, more likely than not that any one of them did not pay. L. J. COHEN, THE PROVABLE AND THE PROBABLE 74–76 (1977).

It is generally conceded that the Blue Bus Company will avoid liability for the accident, even though, based on the statistics, the Company is more likely liable than not.[332]

The Blue Bus case presents the issue of how to consider the different risks of error. The Type I error would be to accept the statistical argument and find the Blue Bus Company liable when it was really innocent. The Type II error would be to overlook the statistical evidence and find the Blue Bus Company not liable when it was really culpable. If we assume that bus accidents are proportional to the number of busses each bus company owns, we can calculate how often we would make each type of error, depending on whether we accept the naked statistical evidence or not. If we would always find the Blue Bus Company liable based on the statistical evidence alone, then we would find it liable one hundred percent of the time, even though it only accounted for eighty percent of the accidents. We would wrongly find liability, the Type I error, twenty percent of the time. If we never found the Blue Bus Company liable based on the statistical evidence, then we would never find it liable, even though it accounted for eighty percent of the accidents. We would wrongly find no liability, the Type II error, eighty percent of the time. By ignoring the statistical evidence, we are saying that we are willing to make four times as many Type II errors as Type I errors.

One of the more common justifications for not finding liability in the Blue Bus scenario is that "the plaintiff's failure to adduce some further evidence appears unjustified, because such evidence should be available to them at little cost."[333] Thus, the lack of non-statistical evidence of guilt is itself evidence of innocence.

But that rationale does not cover what the hypothetical implies—the situation where there is no other evidence to be had.[334] Assume that the accident was not severe enough to cause damage to the bus because the victim was a pedestrian, that it was a rainy evening, which means any blood would have washed off the bus, and that all a review of other evidence revealed was a confirmation that eighty percent of the buses that could have

---

332. In the case from which the Blue Bus hypothetical was drawn, *Smith v. Rapid Transit, Inc.*, 58 N.E.2d 754 (Mass. 1945), a bus company that had the sole franchise for operating buses on the street where an accident occurred escaped liability when there was no evidence of wrongdoing because the court found that private or chartered buses could also have used the street: "The most that can be said of the evidence in the instant case is that perhaps the mathematical chances somewhat favor the proposition that a bus of the defendant caused the accident. This was not enough." *Id.* at 755.

333. D.H. Kaye, *Apples and Oranges: Confidence Coefficients and the Burden of Persuasion*, 73 CORNELL L. REV. 54, 56 (1987). Judge Richard Posner agrees that there should be a negative inference from the lack of supporting evidence, and also contends that the benefit of trying those few cases for which other evidence could not be readily obtained is marginal. Posner, *supra* note 7, at 1508–10.

334. Ronald J. Allen, *A Reconceptualization of Civil Trials*, 66 B.U. L. REV. 401, 412 (1986) ("The only sensible way to understand the hypothetical is that it presents the question of what should be done when this is all the evidence there is.").

caused the accident were owned by the Blue Bus Company. Or, we could consider a case where all of the alternate evidence is destroyed through no fault of either party.[335]

There would still be great reluctance to let the numbers, by themselves, prove liability in such cases. Why? One way to think about this is to consider cases where the numbers do prove liability.[336]

An employer can be found to have violated Title VII of the Civil Rights Act of 1964 based on statistics alone that show that a particular employer's practice had a "disparate impact" based on race or gender.[337] The reason naked statistical evidence suffices for a finding of liability is that Title VII prohibits "not only overt discrimination but also practices that are fair in form, but discriminatory in operation."[338] Numbers are quite capable of communicating that a discriminatory result occurred. The numbers are not being offered into evidence to show a discriminatory intent on the part of the employer.

In extreme cases, however, numbers can indeed be a window into human motivations. For example, in *Yick Wo v. Hopkins*,[339] the Court found a constitutional violation from the fact that more than 200 Chinese laundry owners had been denied permits to operate their business, while 80 non-Chinese owners were granted licenses.[340] According to the Court, this numerical disparity was sufficient to "require the conclusion" that the

335. *See* Craig R. Callen, *Adjudication and the Appearance of Statistical Evidence*, 65 TUL. L. REV. 457, 470 (1991) ("Assuming that all the evidence except one piece has been destroyed without anyone's fault, a factfinder can still wonder whether that one piece of evidence, if weak, is sufficient to support a verdict.") (citations omitted).

336. *See, e.g.*, Callen, *supra* note 335, at 471 ("Courts often hold statistical evidence sufficient to support a verdict. For example, judges rely heavily on statistics in [T]itle VII cases, and no one seriously questions their sufficiency. The question is whether, and under what circumstances, statistics are a sufficient basis for a verdict or for a refusal to enter summary judgment.") (citation omitted).

337. *See, e.g.*, Wards Cove Packing Co. v. Atonio, 490 U.S. 642 (1989); Griggs v. Duke Power Co., 401 U.S. 424 (1971). Title VII of the Civil Rights Act of 1964 states that it is an unfair employment practice for an employer

> (1) to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin; or (2) to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's race, color, religion, sex, or national origin.

Civil Rights Act of 1964, Pub. L. No. 88-352 § 703, 78 Stat. 255 (1964) (codified as amended at 42 U.S.C. § 2000e-2(a)(1)-(2) (2006)).

338. *Griggs*, 401 U.S. at 431.

339. 118 U.S. 356 (1886).

340. *Id.* at 359.

licensing decisions had been made "with a mind so unequal and oppressive as to amount to a practical denial by the state of that equal protection of the laws . . . ."[341]

Similarly, the Supreme Court in *Gomillion v. Lightfoot*[342] concluded that an extreme numerical discrepancy could reveal a bigoted mind. In reviewing a redistricting plan for Tuskegee, Alabama, which removed from the city all but four or five of its 400 African-American voters without removing a single white voter, the Court declared that "the conclusion would be irresistible, *tantamount for all practical purposes to a mathematical demonstration*, that the legislation is solely concerned with segregating white and colored voters by fencing Negro citizens out of town so as to deprive them of their pre-existing municipal vote."[343]

Illegal motivation can also be inferred numerically by, what the Court has termed, "the inexorable zero."[344] Under this doctrine, the fact that an employer had hired no women or minorities when some were arguably available would lead to an inference of discriminatory motive.[345] In a sex discrimination suit against a drug store company, where evidence showed that the company had hired hundreds of male manager trainees and not chosen a single woman during a seven and one-half year period, the Fifth Circuit Court of Appeals rejected the employer's contention that "zero is just a number."[346] The court explained that "zero . . . carries special significance in discerning firm policies and attitudes" because even the hiring of two or three women would indicate "at least some willingness to consider women as equals in firm management."[347] The total absence of such hiring, in contrast, indicates an unwillingness to view women as equals, and accordingly has led courts to be, "particularly dubious of attempts by employers to explain away 'the inexorable zero' when the hiring columns are totalled."[348]

But, absent extreme cases, courts are properly reluctant to find either bad motive or wrongful actions based on naked statistics as the simple luck

---

341. *Id.* at 373.

342. 364 U.S. 339 (1960).

343. *Id.* at 341 (emphasis added).

344. Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 342 n.23 (1977). Ironically, the facts of *International Brotherhood of Teamsters* did not actually involve a "zero," just a large disparity. Of the 2,919 whites who held driving jobs 1,802 (62%) were higher paid line drivers and 1,117 (38%) were city drivers; of 180 African-American and Spanish-surnamed Americans only thirteen (7%) were line drivers and 167 (93%) were city drivers. *Id.*

345. Note, *The "Inexorable Zero,"* 117 HARV. L. REV. 1215, 1215–16 (2004).

346. Capaci v. Katz & Besthoff, Inc., 711 F.2d 647, 662 (5th Cir. 1983).

347. *Id.* In a slightly bizarre analysis, the court distinguished judges from "the noble theoretician predicting the collisions of weightless elephants on frictionless roller skates [for whom] zero may be just another integer . . . ." *Id.*

348. *Id.*

of the draw might well supply the reason for a numerical disparity. In such a case, with no other evidence available, we can treat random chance as a plausible alibi witness. Thus, courts should be particularly concerned with avoiding the Type I error of incorrectly finding an improper motive or action in the absence of other evidence, even if that concern leads to an increase in Type II errors of exonerating the culpable.[349]

Evidence of wrongdoing, however, should lead a court to recalibrate its view of the proper balance between Type I and Type II errors. Imagine that we knew that the Blue Bus Company's drivers had been drinking at a bar before they began their routes that evening.[350] With some evidence of wrongdoing, the no-longer naked statistical evidence begins to look more convincing.

Returning to *Heads You Win*, suppose that before Charles had tossed his sixty heads, he was seen filing the edge of his quarter. Assume because either he is not especially adept at filing or he deliberately made sure that his filing was not particularly extreme, his filing resulted in a coin that landed heads fifty-five percent of the time. If we used a significance level of ninety-five percent to determine whether Charles's sixty heads resulted from random chance, we would be more than twenty-four times as likely to mistakenly clear him when his cheating caused his advantage than we are to wrongly condemn him when his advantage was due to random chance.[351]

But why should we accept that disbalance of risks after we have determined that wrongdoing has occurred? Whether judges admit it or not, such a disbalance of risks represents a value judgment as to the relative evil that would result from the different sorts of error. In non-mathematical cases, such as termination of parental rights or permitting the withdrawal of life-sustaining treatment, courts have had little difficulty recognizing that the

---

349. *See supra* notes 304–11 and accompanying text (discussing Type I and Type II errors).

350. The Gatecrasher story, *see supra* note 331, can be similarly amended:

> Suppose in the gatecrasher hypothetical that the operator of the rodeo testified that a particular defendant did not buy a ticket. He knows this, he asserts, because the defendant looks unusual to the operator, the operator sold all the tickets himself, and he would have remembered such an unusual character.

Allen, *supra* note 334, at 415.

351. Using a 5% significance level, we would reject the null hypothesis if and only if there were sixty-one or more heads or tails (as the probability of seeing an event at least this extreme given a fair coin is 3.52%; by contrast, the probability of seeing an event at least as extreme as sixty heads, 5.69%, is above the 5% threshold). By contrast, the probability of seeing an event at least as extreme as sixty-one heads given a coin which yields heads 55% of the time is 13.52%; this comes from a 13.43% chance of seeing at least sixty-one heads and a .09% chance of seeing at least sixty-one tails. This gives an 86.48% chance of wrongfully clearing Charles, which is about 24.57 times as large as the 3.52% chance computed above.

allocation of the risk of error between two outcomes in a particular case must "reflect properly their relative severity."[352] Courts accordingly have understood that it is their responsibility to make the value judgment as to whether "an incorrect finding of fault would produce consequences as undesirable as the consequences that would be produced by an incorrect finding of *no* fault."[353]

By passively accepting the ninety-five percent standard for significance, though, judges are acquiescing in a value judgment that "the social disutility of wrongful inculpation is many times greater than the social disutility of wrongful acquittal."[354] The calculation of "social disutility," however, is properly for courts to make and should vary with the situation. The selection of the significance level should be made to reflect that judicial evaluation.

Judges must realize that they are free to choose significance levels other than ninety-five percent.[355] As noted in the *Reference Guide on Statistics*, "[a]lthough 95% confidence intervals are used commonly, there is nothing special about 95%."[356] Indeed, as one court noted, "[d]ifferent levels of significance may be appropriate for different types of studies depending on how much risk one is willing to accept that the conclusion reached is wrong."[357]

How much risk we are willing to accept must be based on the values of the legal system. Unlike scientific inquiry, "the law is oriented toward the just resolution of cases rather than truth-finding . . . ."[358] While ascertaining the truth is a deep value of our legal system, it is often an unobtainable goal because "verdicts must be rendered even when information is incomplete . . . ."[359] In the face of incomplete information, errors are inevitable. The legal system has long recognized that justice requires a

---

352. Santosky v. Kramer, 455 U.S. 745, 766 (1982).

353. *Id.* at 788 n.13 (Rehnquist, J., dissenting). Then-Associate Justice Rehnquist referred to such a situation as occurring "when the social disutility of error *in either direction* is roughly equal . . . ." *Id.*

354. Cohen, *Confidence in Probability*, *supra* note 284, at 413–14.

355. Judges must recognize the difference between the seeming neutrality of numerical analysis and the value judgments that may lie beneath the surface: "The mechanical quality of the hypothesis test itself may seem to ensure objectivity, but unless the selection of the significance level is also objective and sensible, this seeming objectivity is illusory." Kaye, *Is Proof of Statistical Significance Relevant?*, *supra* note 238, at 1354.

356. Kaye & Freedman, *Reference Guide on Statistics*, *supra* note 10, at 118.

357. Merrell Dow Pharms., Inc. v. Havner, 953 S.W.2d 706, 722 (Tex. 1997).

358. Alexander Morgan Capron, Daubert *and the Quest for Value-Free "Scientific Knowledge" in the Courtroom*, 30 U. RICH. L. REV. 85, 86 (1996); *see also* Finley, *supra* note 325, at 366 ("The tort system should not slavishly follow the values of epidemiology because its purposes and social functions have always included a 'justice' role that is broader in scope than whether scientists have arrived at a conclusion, or whatever happens to be the scientific 'truth' consensus of the moment.").

359. Capron, *supra* note 358, at 86.

different weighting of risks of error after a finding of wrongdoing by one of the parties.

For example, in numerous cases, courts have shifted the burden of proof of causation away from an innocent plaintiff and onto the negligent defendant, when requiring the plaintiff to prove causation "would be both unfair and destructive of the deterrent purposes embodied in the concept of duty of care."[360] One such case is *Kingston v. Chicago & Northwest Railroad Co.*,[361] in which the plaintiff's property had been damaged when two fires united. One fire was attributable to the negligence of the defendant railroad, the other was of unknown origin.[362] The court held that the defendant should carry the burden of proving that the fire set by him was not the proximate cause of the damage.[363] The reason for this burden-shifting, according to the court, was that forcing the plaintiff to prove the causation in such a situation "would certainly make a wrongdoer a favorite of the law at the expense of an innocent sufferer."[364]

Similar reasoning was employed in *Sindell v. Abbott Laboratories*,[365] a case in which children injured by their mothers' ingestion of DES sued several DES manufacturers, and were unable to identify who manufactured the drugs taken by their mothers. The California Supreme Court held that, even though the plaintiffs could not show which defendant caused their injuries, each defendant would be presumptively liable for a portion of the judgment based on its share of the market.[366] The court gave two reasons for removing the burden of proving causation from the plaintiffs. First, the court said that justice required balancing the risks of uncertainty between the parties: "[A]s between an innocent plaintiff and negligent defendants, the latter should bear the cost of the injury."[367]

---

360. Price Waterhouse v. Hopkins, 490 U.S. 228, 263 (1989) (O'Connor, J., concurring).

361. 211 N.W. 913 (Wis. 1927).

362. *Id.* at 914.

363. *Id.* at 915.

364. *Id.* One commentator has described the holding in *Kingston* as an attempt by the court to devise

> an evidence rule that would balance the competing claims. The court protected individualistic values by retaining the defendant's right to be free from liability unless there was proof of causation. Status concerns were protected by switching to the negligent defendant the burden of establishing the cause of the other fire.

Lawrence W. Kessler, *Alternative Liability in Litigation Malpractice Actions: Eradicating the Last Resort of Scoundrels*, 37 SAN DIEGO L. REV. 401, 460 (2000).

365. 607 P.2d 924 (Cal. 1980).

366. *Id.* at 936.

367. *Id.* at 937.

Second, the court noted that tort law is designed to deter wrongful conduct.[368] In *Sindell*, the DES manufacturers had known that there was a "grave danger" that DES could cause cancer in the daughters of pregnant women who took the drug, yet continued to market the drug as a miscarriage preventative.[369] The manufacturers also failed to test DES for safety and ignored tests performed by others that indicated that the drug was not safe.[370]

In light of such wrongdoing by the defendants, the court ruled that the DES manufacturers should be forced to carry the burden of proof of causation based on a "broader policy standpoint."[371] The DES manufacturers were in "the best position to discover and guard against defects in [their] products and to warn of harmful effects; thus, holding [them] liable for defects and failure to warn of harmful effects will provide an incentive to product safety."[372]

A similar shift in the burden of proving causation has occurred in employment discrimination cases.[373] If an employer is shown to have engaged in a discriminatory hiring pattern and practice, individual employees need not prove that the employer's discrimination was the cause of adverse treatment that the employees personally, individually received.[374] Instead there is a rebuttable presumption that an employee was the victim of the employer's discriminatory practices, and the burden shifts to the employer to overcome that presumption for each employee.[375]

The Supreme Court has explained this burden-shifting in part with the normative rationale that the employer who has committed a pattern of discrimination can no longer be viewed as a presumptively benign actor.[376] It is appropriate to shift the burden of proving individual causation to the employer because, "the finding of a pattern or practice changed the position

368. *Id.*

369. *Id.* at 925.

370. *Id.* at 926.

371. *Id.* at 936.

372. *Id.*

373. A shift in burdens of proof has also occurred in the criminal context. The Supreme Court has held that the prosecution need only prove the factors that go into sentencing decisions by a "preponderance of the evidence." McMillan v. Pennsylvania, 477 U.S. 79, 91–92 (1986). Similarly, in *United States v. Watts*, 519 U.S. 148, 157 (1997), the Court held that an acquittal on the charge of using a firearm during a drug offense, did not preclude the judge, during the sentencing phase, from determining by a preponderance of the evidence that the defendant did, in fact, use such a weapon during a drug offense. As has been discussed, this means that the risk of the defendant suffering an erroneously increased sentence is treated as no more serious than the risk of the defendant enjoying an erroneously decreased sentence. Normally, the law considers the harm of a wrongful conviction as far greater than the harm of an erroneous acquittal. That calculus shifts in the sentencing phase because, "criminal sentencing takes place only after a defendant has been adjudged guilty beyond a reasonable doubt." *McMillan*, 477 U.S at 92 n.8.

374. Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 359 (1977).

375. *Id.* at 359 n.45.

376. *Id.* at 359.

of the employer to that of a proved wrongdoer."[377] On a practical and logical level, that change to the status of "proved wrongdoer" denied the employer the ability to "claim that there is no reason to believe that its individual employment decisions were discriminatorily based, it has already been shown to have maintained a policy of discriminatory decisionmaking."[378] In the case of a proved discriminatory pattern, an employee's detrimental treatment should be presumed to have been caused by discrimination because there was "a greater likelihood that any single decision was a component of the overall pattern."[379]

An analogous burden-shifting should occur in cases involving statistical evidence. The allocation of the risk of error should not be the same when Charles is a "proved wrongdoer," guilty of filing the coin, as when there is no evidence that he committed any wrongdoing. After his attempt to alter the coin, there is undoubtedly "a greater likelihood" that the coin toss was not a fair one.

We cannot directly calculate the effect this increased likelihood of an unfair toss has on the statistical probability previously obtained. In theory, Bayes' Theorem should provide an equation that permits the combination of the probability that Charles changed the coin with the statistical probability into a neat mathematical formula.[380] Unfortunately, Bayes' formula cannot provide us that information because determining the increased probability that the coin tosses were unfair is "necessarily subjective."[381] Because the assessment of that "increased probability" is nothing more than one's imprecise, non-scientific belief as to the likelihood that Charles succeeded in his attempt to alter the coin, there is no number we can plug into the equation to get the correct probability. This is, then, one of those instances where Bayes' Theorem is most useful as a heuristic device, reminding us that the persuasive power of statistical evidence depends, in part, on whether it is consistent with or at variance with the indications we can draw from the other relevant evidence we may have.[382] Taking heed of that lesson, the task

377. *Id.* at 359 n.45; *see also* McKenzie v. Sawyer, 684 F.2d 62, 77 (D.C. Cir. 1982) ("[A]ll doubts are to be resolved against the proven discriminator rather than the innocent employee.").

378. *Int'l Bhd. of Teamsters*, 431 U.S. at 362.

379. *Id.* at 359 n.45; *see also* Davis v. Coca-Cola Bottling Co., 516 F.3d 955, 966 (11th Cir. 2008) ("Because the court's finding of a pattern or practice changed the position of the employer to that of a proved wrongdoer, each class member seeking redress may rely on that finding as circumstantial evidence that the employer made the challenged employment decision with intent to discriminate.") (internal quotation marks omitted).

380. *See supra* notes 69–92 and accompanying text.

381. Kaye & Freedman, *Reference Guide on Statistics*, *supra* note 10, at 127.

382. *See* Allen, *supra* note 334, at 402 ("It is becoming increasingly obvious, for example, that

becomes determining the proper way to combine the hard numbers from the statistical test with the information that the defendant is a proven wrongdoer.

The question can be better approached by considering the issue of securities fraud. Under Rule 10b-5, it is illegal to manipulate stock prices by making false or misleading material statements.[383] In order to prevail on a Rule 10b-5 claim, plaintiffs must show that they relied on those false or misleading statements to their detriment.[384] In *Basic Inc. v. Levinson*,[385] the Supreme Court ruled that a rebuttable presumption of reliance could be created through the economic theory known as "fraud-on-the-market."[386]

According to the Court:

> The fraud on the market theory is based on the hypothesis that, in an open and developed securities market, the price of a company's stock is determined by the available material information regarding the company and its business . . . . Misleading statements will therefore defraud purchasers of stock even if the purchasers do not directly rely on the misstatements . . . .[387]

In other words, once a plaintiff proves the false or misleading statements, the next step is to show that those statements affected the market price of the stock to the plaintiff's detriment.[388]

Generally, economists conduct what is known as an "event study" to compare the actual return on a stock directly after the misleading statement is given with the predicted return, yielding a mathematical estimate of what the return would have been absent such statement.[389] This estimate is based

---

Bayesian approaches can best be used heuristically as guides to rational thought and not as specific blueprints for forensic decisionmaking."); Lempert, *supra* note 330, at 446 ("Bayes's Theorem may be useful as a heuristic device.").

383. 17 C.F.R. § 240.10b-5 (2009) states:

> It shall be unlawful for any person, directly or indirectly, by the use of any means or instrumentality of interstate commerce, or of the mails or of any facility of any national securities exchange . . . .
>
> . . . .
>
> (b) To make any untrue statement of a material fact or to omit to state a material fact necessary in order to make the statements made, in the light of the circumstances under which they were made, not misleading . . . .

384. According to the Supreme Court, "the burden is on the plaintiff to show the violation or the fact that the statement was false or misleading, and that he relied thereon to his damage." Ernst & Ernst v. Hochfelder, 425 U.S. 185, 206 (1976) (quoting S. Rep. No. 792, 73d Cong., 2d Sess., 12–13 (1934)).

385. 485 U.S. 224, 241–42 (1988).

386. *Id.* at 241.

387. *Id.* at 241–42 (quoting Peil v. Speiser, 806 F.2d 1154, 1160–61 (3rd Cir. 1986)).

388. This assumes that it can be shown that the market for the securities in question was "efficient." Macey et al., *supra* note 261, at 1022–28.

389. *Id.* at 1029.

on a statistical test called "regression analysis" which factors in both the firm's average return during some control period as well as any contemporaneous market-wide influences, such as news affecting the entire relevant market which would likely have affected the firm in question as well.[390] The difference between the actual return and the predicted return is called the "abnormal return."[391]

Some variation in returns can be expected, of course, due to the random chance of a volatile market. A large abnormal return, however, indicates that it is unlikely that the market was unaffected by the misleading statement. To determine if the variation in return was caused by random chance, economists use traditional hypothesis testing.[392]

First, they establish a null hypothesis that the misleading statements had no effect on the market.[393] Based on the number of standard deviations that the abnormal return is from the predicted return, they can then calculate the probability of seeing an abnormal return of such magnitude based purely on chance.[394]

Assume that a corporation gives out deliberately misleading information implying high quarterly profits.[395] Stock prices rise, but the company subsequently announces a negative earnings report for that quarter, and the price of the stock plummets. Shareholders who had purchased after the dissemination of the misleading information but before the earnings report was released sue the corporation, and a regression analysis shows that there is only a ten percent or fifteen percent probability of seeing an abnormal return as large as that experienced by the company based purely on chance.

The question of whether those statistics indicate that the market was indeed affected by the misleading statements will turn on our choice of a significance level.[396] According to four of the leading scholars in this area, Jonathan Macey, Geoffrey Miller, Mark Mitchell, and Jeffry Netter, the classic P-value of .05 should be utilized: "We suggest choosing a

390. For an excellent discussion of regression analysis, see Rubinfeld, *supra* note 276, at 1065–68. *See also* John E. Lopatka & William H. Page, *Economic Authority and the Limits of Expertise in Antitrust Cases*, 90 CORNELL L. REV. 617, 688–94 (2005).

391. Macey et al., *supra* note 261, at 1029.

392. *Id.* at 1037.

393. *Id.* at 1040.

394. *Id.*

395. This fact pattern is derived from *Shaw v. Digital Equipment Corp.*, 82 F.3d 1194 (1st Cir. 1996).

396. Macey et al., *supra* note 261, at 1041.

significance level such that the probability of a Type 1 error is less than 5%; this is a standard level used by researchers in finance and economics."[397]

The Type I error with which they are concerned is that of finding an effect on the market when the abnormal return was in fact the result of random chance.[398] Their proposal, however, completely disregards the probability of a Type II error, finding no effect on the market when the misleading statement actually caused harm.[399] As we have seen, the .05 level leads to a decision-making regime in which the probability of an incorrect exoneration far exceeds the probability of an incorrect condemnation.[400]

If we utilized the traditional confidence level of ninety-five percent, the statistician would conclude that the statistics were "not significant" and thus not proof that the misleading information affected the market. But what if a different confidence level had been selected, such as eighty percent?[401] Now, the statistics would be found to be "significant" and proof that the market was affected.

The usual response to a suggestion of an eighty percent significance level is that it would permit too high a rate of error. That concern, of course, only refers to the Type I error of finding an effect on the market when the abnormal return was really the result of random chance. But if our concern is with avoiding too high a rate of Type II error, the ninety-five percent significance level is also suspect. In fact, there is an arbitrariness to choosing any magic number for a significance level, especially when we are concerned with both types of errors.

The ideal solution would be to use a significance test that had the effect of equalizing both Type I and Type II errors. Such an approach would reflect the assumption implicit in the preponderance of the evidence standard for civil trials in general, that Type I and Type II errors impose "essentially equal costs" on society.[402] It would also "equalize the cost of 'wrong'

397. *Id.* The authors do add that "there is no correct significance level, and calibrating the tradeoff is ultimately a value judgment based on the costs of incorrectly rejecting the null hypothesis." *Id.*

398. *Id.*

399. *Id.*

400. *See supra* notes 312–15 and accompanying text.

401. *See* Farrell, *supra* note 323, at 2211. Professor Farrell proposes that, for normative reasons, where the purpose of a legal decision is to award compensation for personal loss, the law should adopt a lower degree of certainty, perhaps an 80% standard of statistical significance, and thus display greater tolerance for false positives. *See id.*

402. Cohen, *The Gatekeeping Role*, *supra* note 238, at 950; *see also* Posner, *supra* note 7, at 1504 ("In the typical civil trial, there is no basis for supposing that Type I errors (false positives, such as convicting an innocent person) on average impose higher costs than Type II errors (false negatives, such as an erroneous acquittal).").

judgments so that the system as a whole would favor neither plaintiffs nor defendants."[403]

One mathematical problem with implementing such a solution is that Type II errors are often not able to be calculated.[404] To understand why, consider this story of a challenge to a grand jury pool.[405] In a community where thirty-eight percent of African-Americans were eligible to serve, eighteen persons were selected by local jury commissioners to serve on the grand jury; three of those selected were African-Americans (seventeen percent) and the other fifteen were white. The question is whether the disparity between the expected percentage and the actual percentage was the result of random chance or a discriminatory selection process.

With the usual null hypothesis of random chance, the P-value for this situation is .051, which would not be "statistically significant" at the traditional .05 level. That .05 level, remember, represents a five percent risk of the Type I error of incorrectly condemning the innocent.

Calculating the risk of the Type II error of incorrectly exonerating discriminatory jury commissioners is impossible, because we do not know how discriminatory they were. As one commentator noted, "there are many possible degrees of inequality or disadvantage."[406] The commissioners could be enthusiastic bigots, trying to prevent every black from serving, or they could be subtler bigots, just tilting the playing field slightly. Intuitively, it is easier to mistakenly overlook the discriminator who only marginally affects the selection process as opposed to the one whose results are far more blatant.

Statisticians capture that variability with what is known as the "power function."[407] The power function represents the differing probabilities of rejecting the null hypothesis of "no effect" for the full range of possible actual effects. In the grand jury case, for example, the power function would

---

403. Cohen, *Confidence in Probability*, *supra* note 284, at 417; *see also* Dawson, *supra* note 315, at 209.

404. *See* FORTHOFER, *supra* note 306, at 218.

405. This story is derived from Kaye, *Is Proof of Statistical Significance Relevant?*, *supra* note 238, at 1338–60. Professor Kaye's example was modeled after *Moultrie v. Martin*, 690 F.2d 1078, 1082 (4th Cir. 1982).

406. *See* Goldstein, *supra* note 311, at 35. As Dr. Goldstein notes:

> [T]he true rate at which one group is being hired may be six percent higher than another (a simple alternative hypothesis), or it may be four percent higher (another simple alternative), or five, or seven, or eight percent higher, and so on. The vaguer alternative hypothesis of some degree of difference or disadvantage is a composite of such simple, well-specified, hypotheses.

Id. at 35–36.

407. Kaye, *Is Proof of Statistical Significance Relevant?*, *supra* note 238, at 1357.

reveal that use of the ninety-five percent confidence level will very rarely identify a small discriminatory effect. The test has a somewhat improved chance of identifying larger discriminatory effects but still "does not have a better than even chance of correctly detecting [discrimination]—unless the list is so grossly biased . . . that a black's chance of appearing on a grand jury is diluted by some sixty percent" from what it would be absent any discrimination.[408]

It is obvious in the grand jury example that use of the ninety-five percent significance level will create a high risk of overlooking discriminatory conduct. We cannot simply say "equalize" Type I and Type II errors, however, because there are multiple probabilities for different Type II errors.

One way to fulfill the spirit of equalizing Type I and Type II errors is to borrow from the concept known as "baseball" or "final offer" arbitration.[409] In baseball arbitration, two parties each submit an amount meant to represent a fair resolution of a dispute and the arbitrator must choose between those two. Thus, in a salary dispute, the team owner and player each select a proposed salary and the arbitrator must select one or the other as the "fairer" figure.

For hypothesis testing, each side could propose their own hypothesis. Thus, the jury commissioners charged with discrimination would provide the null hypothesis; they would then contend that their conduct had "no effect" on the actual numbers that appeared. Those challenging the commissioners would need to select some level of discrimination they contend occurred as the alternate hypothesis. From those two possibilities, it is easy to create a significance level that equalizes Type I and Type II errors. If the regression analysis revealed a higher probability of seeing an abnormal return than that significance level, the statistics would not disprove the defendant's hypothesis that its statements did not affect the market. If however, the analysis revealed a smaller probability of seeing that abnormal return, the plaintiff's alternate hypothesis with the specified level of discrimination would be accepted.[410]

The "baseball arbitration" approach would be particularly valuable in cases such as the securities fraud example discussed earlier.[411] A

---

408. *Id.* at 1359.

409. Keith Sharfman, *Valuation Averaging: A New Procedure for Resolving Valuation Disputes*, 88 MINN. L. REV. 357, 365–66 (2003).

410. Even if a judge were reluctant to alter the usual .05 significance level, such an equalization of risks would be appropriate if there was some other evidence of discriminatory actions by the jury commissioners. For example, as the court in the case on which this example was drawn stated, "[a]dditional evidence that could have supported the petitioner's case would include statistics showing that the jury commissioners exempted a disproportionate number of blacks during their proceedings." *Moultrie*, 690 F.2d at 1085.

411. *See supra* text accompanying notes 383–89 (discussing an example of securities fraud).

corporation that had deliberately released misleading information, that is, a proven wrongdoer, would contend for its null hypothesis that its statements had no effect on stock prices.[412] For their alternate hypothesis, plaintiff shareholders would then select some level of price increase that they would claim was caused by the misleading statements for their alternate hypothesis. Again, from those two hypotheses, a significance level could be selected to balance the risk of mistakenly finding an effect from a harmless statement and erroneously excusing the company whose fraud affected the market. Note that the smaller the alleged effect, the lower the equalizing significance level will need to be, making it easier for the plaintiff to prove an effect. Of course, the smaller the alleged effect, the lower the plaintiff's damages will be. To prevent the manipulation of statistics with *de minimis* alleged effects, courts could require that any effect be of "practical or substantive significance."[413]

Equalizing Type I and Type II errors would do more than simply balance the risks of error. It would also reflect a policy judgment that the harm from the risk of error to an innocent investor is deemed as important as the harm of the risk of error to a proven wrongdoer.

Sometimes, however, the nature of the statistical analysis being used makes it difficult to create a power function, and thus impossible to equalize the risk of error.[414] Additionally, there may be judges who understand the necessity of improving the balance between Type I and Type II errors but still feel wedded to the ninety-five percent confidence level. For both situations, there is another approach: When dealing with a proven wrongdoer, courts should utilize the "one-tailed" rather than the more traditional "two-tailed" analysis.[415]

To understand the distinction, recall that hypothesis testing determines the probability of obtaining a result as "extreme" as the one actually seen if the null hypothesis of "no effect" were correct.[416] In *Heads You Win*, the two-tailed test revealed how likely it was for Charles to toss sixty or more

412. There might be strategic reasons for the company to select some small level of price increase, for example, if the use of that small increase creates a better chance of the null hypothesis not being disproved.

413. Goldstein, *supra* note 311, at 38 n.13.

414. *See, e.g.*, *id.* at 36 n.10 ("For some forms of statistical analysis, the calculation of power is very complicated and/or involves a noncentrality parameter (measure of the extent of difference in the populations) that has no obvious or intuitive interpretation.").

415. Stender v. Lucky Stores, 803 F. Supp. 259, 323 (N.D. Cal. 1992).

416. *See supra* text accompanying notes 281–83.

heads *or* sixty or more tails. A one-tailed test would focus on the narrower question of how likely it was for him to toss sixty or more heads.

On a strictly numerical basis, with a one-tailed test it is usually easier to find statistical significance. A one-tailed test with a significant value of ninety-five percent is generally equivalent to a two-tailed test with a significant value of ninety percent.[417] That would mean that Type I errors would be twice as likely, and Type II errors would be less likely (though by an uncertain amount).[418] Thus, Type I and Type II errors would be more nearly in balance; there is a greater chance of condemning the innocent and a lower risk of exonerating the culpable.[419]

In the case of a proven wrongdoer, this improved balance is important. Some courts, however, have rejected one-tailed tests precisely because they make it easier for plaintiffs to show a statistical significance.[420] One court in an employment discrimination case explained its preference for the two-tailed test on the grounds that a one-tailed test only indicates whether blacks are treated worse than whites while a two-tailed test "demonstrates whether blacks and whites were treated equally, taking into account both whether whites are treated as well as or better than blacks and vice versa."[421]

That argument only makes sense, though, if there is reason to believe that the discrimination could have gone in either direction. If there is other evidence of anti-black bias by an employer, it would be naïve at best to pretend that whites are as likely to be harmed as African-Americans. For example, the Supreme Court has stated that where a school board acted with a discriminatory motive, there is a high probability that "similar impermissible considerations have motivated their actions in other areas of the system."[422] With evidence of discriminatory intent, a one-tailed test is a far more accurate tool for assessing how likely the "extreme" result was to

---

417. The practical difference between one-tailed tests and two-tailed tests "is that the P-value produced by a two-tailed test is usually twice as great as that produced by a one-tailed test." *Stender*, 803 F. Supp. at 323, 323.

418. As Dr. Richard Goldstein wrote, "[a] one-tailed test is more powerful." Goldstein, *supra* note 311, at 47.

419. *See supra* text accompanying notes 235–37. While we know that the probability of a Type II error will decrease with a one-tailed analysis, it is impossible to say by how much. *See supra* notes 232–33 and accompanying text. Also, even after the decrease, there still would likely be a larger probability of a Type II error than a Type I error.

420. *See* Palmer v. Shultz, 815 F.2d 84, 96 (D.C. Cir. 1987) (stating that "a two-tailed test and a 5% probability of randomness require statistical evidence measuring 1.96 standard deviations. Consequently, if plaintiffs come into court relying *only* on evidence that the underselection of women for a particular job measured 1.75 standard deviations, it seems improper for a court to establish an inference of disparate treatment on the basis of this evidence alone.").

421. Moore v. Summers, 113 F. Supp. 2d 5, 20 n.2 (D.D.C. 2000).

422. Keyes v. Sch. Dist., 413 U.S. 189, 208 (1973).

occur: "One-tailed tests are most appropriate when one population is consistently overselected over another."[423]

Courts should be able to act on the reasonable presumption that the proven wrongdoer who attempted to create an effect in one direction did not create an effect in the opposite direction. Thus, if Charles filed the coin in order to increase the number of heads he tossed, we should not assume he was equally as likely to increase the numbers of tails he tossed. Similarly, in the case of the securities fraud, if we determine that the deliberately misleading statements were made in an attempt to raise stock prices, it would be perfectly reasonable to assume they did not lower prices.

Although conceivable, it is highly unlikely that a truly incompetent wrongdoer was completely counterproductive. When courts use a two-tailed test, they are essentially giving an unjust benefit of the doubt to the party who was actively committing wrongful acts.[424] Use of the two-tailed test in such a situation "would certainly make a wrongdoer a favorite of the law at the expense of an innocent sufferer."[425] In the name of justice, as well as probability, courts should be willing to analyze cases of a proven wrongdoer by use of a one-tailed test.

VI. CONCLUSION

More than forty years ago, the California Supreme Court admonished: "Mathematics, a veritable sorcerer in our computerized society, while assisting the trier of fact in the search for truth, must not cast a spell over him."[426] It turns out that mathematics is not merely a sorcerer but a bully as well, seizing the power to make policy judgments that belong to the courts.

In the misguided name of mathematical rigor, courts have allowed "race talk" to enter into our criminal trials, and prejudice to reduce tort awards. With judicial acquiescence, information universally acknowledged to be incorrect and baseless has been used to calculate the probabilities of paternity. Moreover, known wrongdoers have been permitted to escape liability because judges have abdicated their responsibility of balancing the

423. Stender v. Lucky Stores, 803 F. Supp. 259, 323 (N.D. Cal. 1992); *see also* United States v. Delaware, 2004 U.S. Dist. LEXIS 4560, at *36 n.27 (D. Del. Mar. 22, 2004) (stating that for a one-tailed test to be appropriate, "one must assume . . . that there will only be one type of relationship between the variables").

424. Kingston v. Chicago & Nw. R.R. Co., 211 N.W. 913, 915 (Wis. 1927).

425. *Id.* at 915.

426. People v. Collins, 438 P.2d 33, 33 (Cal. 1968).

risks of error that are an inevitable part of any trial. Outside the realm of mathematics, judges would never accept such policies.

It is not necessary for judges to become "amateur mathematicians" in order to reclaim their rightful role.[427] However, they must be aware that the apparent objectivity of mathematics often masks subjective judgments, and not be fooled when "hard" numbers are really based on little more than intuition and guesswork.[428] Numbers can communicate important information. Judges just need to make sure that they are able to comprehend what those numbers are trying to say.

427. In *Daubert*, then-Chief Justice Rehnquist complained that a requirement that judges evaluate "scientific validity," imposes on them the obligation, "to become amateur scientists." Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579, 600–01 (1993) (Rehnquist, C.J., concurring in part and dissenting in part).

428. Kaye, *Is Proof of Statistical Significance Relevant?*, *supra* note 238, at 1347.