4-3-2019

# A Comment on Statistical Evidence and Standards of Proof

Michael S. Pardo
*University of Alabama - School of Law*, mpardo@law.ua.edu

# A Comment on Statistical Evidence and Standards of Proof

Michael S. Pardo

25 SUPREME COURT ECONOMIC REVIEW 59 (2018)

# A Comment on Statistical Significance and Standards of Proof

Michael S. Pardo*

It is a pleasure to have the opportunity to comment on *Error Costs, Legal Standards of Proof, and Statistical Significance*, by Michelle Burtis, Jonah Gelbach, and Bruce Kobayashi. The article provides a clear, sophisticated, and persuasive discussion of the relationships between conventional tests of statistical significance and legal standards of proof. Although tests of statistical significance and legal standards of proof are, in theory, each concerned with similar considerations regarding the types and frequency of inferential errors, the article in my view is correct to argue that they are analytically distinct and that one cannot be mapped onto the other in a straightforward manner. Thus, I agree with their general conclusion: "there is no one level of statistical significance that generally corresponds to the legal standard of proof."[1] The article's related discussion of different types of statistical tests is also persuasive. I thus also agree with their general conclusion that likelihood-ratio tests (comparing competing explanations supporting each side) map onto legal standards of proof more closely than conventional tests with a fixed significance level.[2]

* Henry Upson Sims Professor of Law, University of Alabama School of Law.

[1] Burtis, Gelbach & Kobayashi [manuscript at 3].

[2] Id. [at 5] ("reconciling legal standards of proof and statistical thresholds can be achieved by replacing fixed significance levels with likelihood ratio tests"). In addition to the general conclusions, I found nearly all of the analysis to be persuasive as well, taking issue with only minor details—e.g., the assumption that the "presumption of innocence" in criminal cases refers to prior odds of guilt. Id. [at 17]. For a contrary view, see Larry Laudan, Truth, Error, and Criminal Law: An Essay in Legal Epistemology 90-109 (2006).

Despite the fact that courts, litigants, and academics have sometimes argued or assumed that evidence in the form of conventional tests of statistical significance (at the typical .05 level, or at some other level) is necessary or sufficient to satisfy a legal standard of proof, the article illustrates clearly the problems with these arguments and assumptions. Here is the crux of the analysis: conventional tests of statistical significance typically focus on one type of error (false positives) and are less concerned with false negatives. In other words, the thumb is on the scale of not declaring that a relationship exists (e.g., causation or discrimination) unless it is warranted or justified, but there is less concern with failing to declare a relationship when one does in fact exist. Legal standards of proof, by contrast, are concerned with both types of errors and their costs. The article demonstrates how likelihood-ratio tests that focus on both types of errors are likely to improve accuracy and reduce overall error costs when compared with tests based on fixed significance levels.[3]

In this comment, my aim is to situate the article's analysis in the academic literature and debates on legal standards of proof (focusing on the "preponderance of the evidence" standard). Here is a brief, somewhat simplified, picture of standards of proof and their underlying rationales.[4] Standards of proof focus on the related goals of accuracy and allocating the risk of error between the parties. Common assumptions about the preponderance standard are that it aims at minimizing total errors and allocating the risk

[3] The article illustrates the differences between the tests with an employment-discrimination example. Their analysis of the example should be required reading for any courts, litigants, or academics arguing that there is a clear relationship between statistical significance and legal standards of proof.

[4] The assumptions in this picture are each contested but they are nevertheless common. For a more detailed discussion, see Ronald J. Allen et al., An Analytical Approach to Evidence 803-59 (6th ed. 2016).

of error roughly evenly among civil litigants. These assumptions are justified, in part, by the further assumption that the costs of each type of error will be roughly similar (and thus reducing total errors is likely to reduce to error costs). Moreover, equalizing the risk of error reflects a principle of equality among civil litigants.[5] Asymmetric error costs then justify higher proof standards (e.g., "clear and convincing evidence" and "beyond a reasonable doubt"), which attempt to skew the risk of error away from false positives.

This simple picture gives rise to several distinct theoretical issues, which have generated disagreements among evidence scholars. I will discuss three such issues as they relate to the analysis in the article.

1. *To what extent are the standards comparative?* Standards of proof are sometimes interpreted as probabilistic thresholds (e.g., greater than 0.5) and sometimes as comparative assessments (e.g., a likelihood ratio greater than 1). These interpretations are not the same and they imply different outcomes. Suppose a plaintiff's theory of what happened is 0.4 probable and the defendant's alternative theory is 0.2 probable. Has the plaintiff proven its case by a preponderance of the evidence? The issue concerns what to do with the unknown probability space. Should all the unknown possibilities go against the plaintiff, or should they be divided evenly among the parties? A number of scholars (myself included) have argued, from different perspectives, that a comparative account better explains the legal standards and better fits with their assumed goals regarding

[5] See Mike Redmayne, Standards of Proof in Civil Litigation, 62 Mod. L. Rev. 167, 171-74 (1999); Lawrence B. Solum, Procedural Justice, 78 S. Cal. L. Rev. 181, 286-89 (2004).

accuracy and the risk of error.[6] Although the article focuses on general defense explanations (e.g., "no discrimination"), the likelihood-ratio framework appears to likewise embrace a comparative interpretation of legal standards. One question the article leaves open, however, is how the analysis would change in a situation in which the plaintiff and the defendant are each offering more specific, alternative explanations? In other words, suppose the defendant's explanation is not the null hypothesis (as the article assumes) but instead involves a more specific explanation for their actions. This relates to the distinction the article raises between simple and composite explanations[7], but not necessarily. Neither side may be offering a composite of alternatives—there may just be two simple theories that do not fill the entire space of possibilities.

2. *What are the criteria that underlie the standards?* There is substantial debate on whether the thresholds employed by legal standards of proof are probabilistic thresholds or whether they depend on other criteria (e.g., explanatory thresholds).[8] The article assumes they are probabilistic thresholds of some sort (expressed by either fixed probability thresholds or likelihood ratios), but there is a wrinkle with this assumption, with potentially deep consequences. Legal standards of proof apply to individual elements of claims, not to claims as a whole, so even proving a plaintiff's claim on

[6] See, e.g., Michael S. Pardo & Ronald J. Allen, Juridical Proof and the Best Explanation, 27 Law & Philosophy 223 (2008); Edward K. Cheng, Reconceptualizing the Burden of Proof, 122 Yale L.J. 1254 (2013); Kevin M. Clermont, Standards of Decision in Law 149 (2013).

[7] Burtis, Gelbach & Kobayashi [manuscript at 24].

[8] See Pardo & Allen, supra note 6; Ronald J. Allen & Michael S. Pardo, Relative Plausibility and Its Critics (in progress).

particular elements may create suboptimal results in terms of errors and errors costs.[9] For example, in a two-element claim, *A* and *B*, a plaintiff will win under the preponderance standard by proving each element to 0.6. But if *A* and *B* are probabilistically independent, then the plaintiff's claim is only 0.36 probable. The effect gets worse with more elements.[10] Once again, this is related to—but not quite the same as—the issue of simple versus composite explanations that the authors note.[11] Even simple hypotheses offered by a party may contain multiple legal elements (e.g., "I suffered an adverse employment action because of race" or "the defendant's product caused my injuries"). In sum, the analysis in the article applies most clearly to legal disputes that involve one legal element and one contested factual issue.

3. *What is the relationship between items of evidence and standards of proof*? More generally, the article well illustrates the need to keep separate, as a conceptual matter, evidence, on one hand, and the standard of proof, on the other. The different types of statistical tests examined (those based on fixed significance levels and likelihood ratios) are essentially different types of evidence, and the article provides compelling reasons why the tests are likely to differ in terms of their probative value in proving contested factual issues. As a general matter, the relationship between *any* item of evidence and the

[9] See Ronald J. Allen & Sarah A. Jehl, Burdens of Persuasion in Civil Cases: Algorithms vs. Explanations, 2003 Mich. St. L. Rev. 893; L. Jonathan Cohen, The Probable and The Provable 58-67 (1977)

[10] Probabilistic dependence among elements creates similar problems, see Allen & Pardo, supra note 8. The conjunction effect also applies to comparative probabilistic standards. For example, suppose that in a two-element claim the plaintiff proves one to 0.9 and the other to 0.4, and the probability of the defendant's alternative theory on the elements is 0.1 and 0.6. The plaintiff will lose despite offering a theory that is six times more likely than the alternative (0.36 versus 0.06).

[11] Burtis, Gelbach & Kobayashi [manuscript at 24].

standard of proof is complex, and the probative value of evidence will be defeasible (depending on the other evidence, the specific context, and the contrasting claims and arguments of the parties). This is so for statistical tests regarding discrimination, for relative-risk analysis in proving causation in toxic-tort cases under the preponderance standard, and for random-match probabilities in proving identity in criminal cases under the beyond-a-reasonable-doubt standard, to name just a few vexing examples. This is also true for eyewitness testimony, confessions, or any other non-statistical evidence as well. Thus, one lesson to take from the article's analysis is that various doctrinal rules of thumb (e.g., the 80 percent rule in measuring disparate impact[12] or requiring a relative risk of 2.0 or greater in tort cases[13]), or various presumptions that shift a burden of production or persuasion based on particular items of evidence[14], will always be imperfect guides to legal standards of proof rather than capturing something essential about the relationship between the evidence and the standard of proof.

[12] Id. [at 6].

[13] See Steve C. Gold, The "Reshapement" of the False Negative Asymmetry in Toxic Tort Causation, 37 Wm Mitch. L. Rev. 1507, 1523 (2011) ("courts have equated more than a doubling of relative risk in an exposed group to a more-likely-than-not probability of causation in an exposed individual plaintiff.").

[14] See Allen et al., supra note 4, at 821-35, 857-59.