University of Pennsylvania Carey Law School

Penn Law: Legal Scholarship Repository

Faculty Scholarship at Penn Law

4-21-2017

Error Costs, Legal Standards of Proof and Statistical Significance

Michelle Burtis

Charles River Associates (CRA) International

Jonah B. Gelbach

University of Pennsylvania Carey Law School

Bruce H. Kobayashi

George Mason University School of Law

Follow this and additional works at: https://scholarship.law.upenn.edu/faculty_scholarship

Part of the [Civil Procedure Commons](#), [Courts Commons](#), [Evidence Commons](#), [Law and Economics Commons](#), [Litigation Commons](#), and the [Statistical Methodology Commons](#)

Repository Citation

Burtis, Michelle; Gelbach, Jonah B.; and Kobayashi, Bruce H., "Error Costs, Legal Standards of Proof and Statistical Significance" (2017). *Faculty Scholarship at Penn Law*. 1754.

https://scholarship.law.upenn.edu/faculty_scholarship/1754

This Article is brought to you for free and open access by Penn Law: Legal Scholarship Repository. It has been accepted for inclusion in Faculty Scholarship at Penn Law by an authorized administrator of Penn Law: Legal Scholarship Repository. For more information, please contact PennlawIR@law.upenn.edu.

ERROR COSTS, LEGAL STANDARDS OF PROOF AND STATISTICAL SIGNIFICANCE*

By

Michelle M. Burtis
Vice President
CRA International
mburtis@crai.com

Jonah B. Gelbach
Professor of Law
University of Pennsylvania Law School
jgelbach@law.upenn.edu

Bruce H. Kobayashi
Associate Dean for Research and Faculty Development & Professor of Law
Antonin Scalia Law School at George Mason University
bkobayas@gmu.edu

Version 10.0
April 30, 2017

Abstract

The relationship between legal standards of proof and thresholds of statistical significance is a well-known and studied phenomena in the academic literature. Moreover, the distinction between the two has been recognized in law. For example, in *Matrix v. Siracusano*, the Court unanimously rejected the petitioner's argument that the issue of materiality in a securities class action can be defined by the presence or absence of a statistically significant effect. However, in other contexts, thresholds based on fixed significance levels imported from academic settings continue to be used as a legal standard of proof. Our positive analysis demonstrates how a choice of either a statistical significance threshold or a legal standard of proof represent alternative and often inconsistent attempts to balance error costs, and that thresholds based on fixed significance levels generally are not consistent with existing or optimal legal standards of proof. We also show how the statistical testing and legal standards of proof can be reconciled by replacing fixed significance level hypothesis testing with likelihood ratio tests.

* The authors would like to thank Ron Allen, Jon Klick, Murat Mungan, Jeff Parker, Steve Yelderman, and participants at the University of Michigan Law School's Law & Economics Colloquium for helpful comments on an earlier draft. The opinions expressed are the author's and do not reflect the views of CRA or any of its respective affiliates.

I. Introduction

A primary issue for courts faced with evaluating statistical evidence offered by experts in litigation is how to draw legal inferences from statistical results. One approach taken by some courts is to use the standard tools of *statistical inference* to make *legal inferences*. For example, in a discrimination dispute, the existence of a significant statistical disparity among workers in hiring and salaries, after accounting for other factors, has been used to determine whether the plaintiff has met their burden of production for a *prima facie* case of disparate impact discrimination.¹ Similarly, courts use evidence of statistically significant abnormal returns at the time of a corrective disclosure to determine the existence or absence of price impact in class action securities litigation.²

The principal question addressed in this paper is when and under what circumstances is the use of standard tools of statistical inference consistent with the applicable legal standard of proof that applies to the case.³ Economists, statisticians, and judges import from research settings fixed significance levels—usually 5%, but sometimes 1% or 10%—and use them to make legal inferences.⁴ When a research standard such as the 5% significance level is used to make legal inferences, is the resulting decision consistent with satisfaction of a legal standard of proof such as “more likely than not?” Moreover, if levels of statistical significance used in the research context do not generate results that

¹ See, e.g., *Ricci v. DeStefano*, 557 U.S. 557, 587 (2009) where a showing a statistically significant difference is *prima facie* evidence of “disparate impact” under U.S.C. § 2000e-2(k)(1)(A)(i).

² See *Halliburton Co. v. Erica P. John Fund, Inc.*, 134 S. Ct. 2398, 573 U.S. __ (2014) (*Halliburton II*) (defendant may present evidence of non-significant abnormal returns at class certification stage to show the lack of price impact). See also Jill E. Fisch, Jonah B. Gelbach, and Jonathan Klick, *The Logic and Limits of Event Studies in Securities Fraud Litigation* (February 12, 2017), Texas Law Review, 2018, Forthcoming; ECGI - Law Working Paper No. 328/2016; U of Penn, Inst for Law & Econ Research Paper No. 16-16, <https://ssrn.com/abstract=2817090>; Alon Brav & J. B. Heaton, *Event Studies in Securities Litigation: Low Power, Confounding Effects, and Bias*, 93 WASH. U. L. REV. 583 (2015).

³ The process of fixed significance level null hypothesis testing is described in detail in Section II, below. See also Stephen T. Ziliak & Deirdre N. McCloskey, *THE CULT OF STATISTICAL SIGNIFICANCE: HOW THE STANDARD ERROR COSTS US JOBS, JUSTICE, AND LIVES* (2008) (documenting the arbitrary nature of the standard fixed significance level test, and criticizing its use generally, including uses in academics, science, and law).

⁴ See for example, James H. Stock & Mark W. Watson, *INTRODUCTION TO ECONOMETRICS*, __ ed. (2011) at 77-8 (“In many cases, statisticians and econometricians use a 5% significance level.”); Jeffrey M. Wooldridge, *INTRODUCTORY ECONOMETRICS*, __ ed. (2002) at 124 (“Suppose we have decided on a 5% significance level, as this is the most popular choice. See also Ziliak & McCloskey, *supra* note 3 (discussing the history of significance tests).

are consistent with the applicable legal standard of proof, is there another fixed significance level that does generate such consistent results?⁵

The short answer to both prior questions is no. While both legal standards of proof and statistical significance levels can be characterized as normative choices made to minimize the sum of error costs and direct costs, there is no *a priori* fixed level of statistical significance that will generally coincide with the applicable legal standard of proof. Moreover, it is generally not possible to simply replace one fixed level of statistical significance with other fixed levels of significance to make the statistical test coincide with the applicable legal standard of proof. That is, even though one can express the applicable legal standard of proof in such a way that it can be understood in fixed-significance level terms in each case, the level of statistical significance necessary to achieve this result varies across cases and circumstances. Thus, there is no one level of statistical significance that generally corresponds to the legal standard of proof in a context-free way.

Others have pointed out the divergence between statistical and legal standards of proof. Indeed, there is a large literature examining the distinction between legal standards of proof and statistical significance thresholds.⁶ Moreover, the distinction

⁵ Some courts have considered experts' use of alternative fixed significance levels. See *In Re: High-Tech Employee Antitrust Litigation*, Case No. 11-CV-025009-LHK, Order Re: Defendant's Motions Regarding Dr. Leamer and Defendant's Joint Motion for Summary Judgment Based on Motion to Exclude Testimony of Dr. Leamer, (2014) at 22, fn. 25. In that case, the plaintiffs' expert argued that a 50% significance level "suggests that it is more likely than not that the compensation of employees were decreased during the period of the agreements." The court in *In Re Photochromic Lens Antitrust Litigation*, MDL Docket No. 2173 (04/03/14), also considered statistical evidence offered by plaintiffs that did not meet the "conventional 5% significance level." Plaintiffs argued that a significance level of 50% would "avoid false negatives, or what statisticians refer to as 'Type 2 errors.'" (*id.* at 48). The court in *Photochromic* did not accept the defendants' arguments regarding the impropriety of using a higher fixed significance level, noting that "Although his studies test the boundaries of reliable evidence permitted under *Daubert*, as well as the Supreme Court's directive in *Comcast* that statistical models prove with precision impact and damages on a classwide basis, I cannot agree that Dr. Singer's use of a 50% measure of statistical significance, by itself, is sufficient justification for denying class certification."

⁶ See, e.g., Michael O. Finkelstein and Bruce Levin, *STATISTICS FOR LAWYERS*, 2d. Ed. (2001); Jennifer L. Peresie, *Toward a Coherent Test for Disparate Impact Discrimination*, 84 IND. L. J. 773 (2009) (analyzing differences between EEOC 80 percent rule and statistical significance); Marcel C. Garaud, *Legal Standards and Statistical Proof in Title VII Litigation: In Search of a Coherent Disparate Impact Model*, 139 U. PENN. L. REV. 455 (1990) (discussing the distinction between legal and statistical significance); Neil B. Cohen, *Confidence in Probability: Burdens of Persuasion in a World of Imperfect Knowledge*, 60 N. Y. U. L. REV. 385 (1985) (using error cost analysis to determine appropriate level of confidence in civil litigation); Allan G. King, "Gross Statistical Disparities" as Evidence of a Pattern and Practice of Discrimination: Statistical versus Legal Significance, 22 THE LABOR LAWYER 271 (2007), Richard O. Lempert, *The Significance of Statistical Significance: Two Authors Restate An Incontrovertible Caution. Why A Book?* 14 L. & SOC. INQ. 225 (2009).

between statistical significance thresholds and legal standards of proof has been explicitly recognized by the courts. For example, in *Matrixx v. Siracusano*,⁷ the Court unanimously rejected the petitioner/defendant's argument that equated the materiality standard in a securities fraud case with evidence of "statistically significant" adverse effect of its product on the rate of anosmia (the loss of smell).⁸ Courts also have criticized the use of fixed significance thresholds in antitrust cases.⁹ However, in other contexts, courts continue to use standard fixed-significance level hypothesis testing as a legal standard of proof. For example, in disparate impact discrimination cases, courts continue to use fixed significance level thresholds based on the 5% two-tailed test used by academics in research contexts as the legal standard of proof to prove a *prima facie* case of discrimination "by a preponderance of the evidence."¹⁰ Securities litigation constitutes another example; for example, on remand following *Halliburton II*, the trial court used

⁷ *Matrixx Initiatives, Inc. v. Siracusano*, 563 U.S. 27 (2011). The Court held that available information could indicate a reliable causal relationship between two variables even if that relationship was not statistically significant. In that case, Matrixx, who sold an over-the-counter cold remedy, had information that the remedy was associated with anosmia, the loss of smell. The issue before the Court was whether Matrixx was required to disclose information regarding a relationship between the use of its product and anosmia. Matrixx contended it did not have such an obligation because the information available to it did not indicate a statistically significant relationship. However, the Court found that the lack of a statistically significant relationship did not necessarily mean that there was not a reliable causal relationship between its product and anosmia and that Matrixx's failure to disclose the information was an omission of a material fact.

⁸ See also Joseph L. Gastwirth, *Statistical Considerations Support the Supreme Court's Decision in Matrixx Initiatives v. Siracusano*, 52 JURIMETRICS J. 155 (2012).

⁹ The court in *Photochromic*, *supra* note 5, found that "[t]here is not, however, any 'precise level in the law' at which statistical significance is sufficient to permit the inference derived from a correlative study. And most courts have rejected the arbitrary application of a 5% threshold." (citations omitted).

¹⁰ Garaud, *supra* note 6 at 467. For a recent example, see, e.g., *Jones v. City of Boston*, 752 F.3d 38 (1st Cir. 2014). See also *Alison Palmer v. George P. Shultz*, 815 F.2d 84 (D.C. Cir. 1987) (*prima facie* evidence of disparate impact under Title VII required differences in selection rates that were statistically significant at a 5% level). Numerous other Title VII cases have used statistically significant difference in outcomes as evidence of disparate impact and discrimination. See King, *supra* note 6 at 277 (describing that many lower courts have adopted the criterion of two or three standard deviations which can correspond to a .05 statistical significance level as a bright-line rule). See also *Bennett v. Total Minatome Corp.*, ___ F. ___, *Segar*, 738 F.2d at 1282; *Vuyanich v. Rep. Nat'l Bank of Dallas*, 505 F. Supp. 224, 348 (N.D. Tex. 1980), vacated on other grounds 723 F.2d 1195 (5th Cir. 1984); *Cooper v. Univ. of Tex. at Dallas*, 482 F. Supp. 187, 194 (N.D. Tex. 1979) ("It has become a convention in social science to accept as statistically significant values which have a probability of occurring by chance 5% of the time or less."). See also *Castaneda v. Partida*, 430 U.S. 482, 496 n.17 (1977) ("As a general rule . . . if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the jury drawing was random would be suspect to a social scientist.").

statistical significance as the basis for determining dates on which it was appropriate to certify a Rule 23(b)(3) class action.¹¹

This paper provides an analysis of the statistical significance tests and their relationship to the applicable legal standard of proof, and attempts to clarify the consequences of a choice of statistical significance levels in such settings. Economic analyses characterize legal standards of proof as relative comparisons between competing hypotheses put forward by the litigating parties. The positive analysis demonstrates how a legal standard of proof and statistical hypothesis testing using a fixed statistical significance level represent alternative and generally inconsistent ways to allocate error costs. We also show how reconciling legal standards of proof and statistical thresholds can be achieved by replacing fixed significance levels with likelihood ratio tests that compares the relative properties of the sampling distributions of the competing hypotheses put forward by the parties.¹²

The organization of the paper is as follows. Section II sets out the error cost analysis framework as well as the derivation of optimal legal standards of proof. The analysis then examines the conditions under which existing legal standards, such as the preponderance rule, are optimal standards. In Section III we describe the standard use of null hypothesis testing under fixed significance levels. Section IV examines the relationship between statistical significance levels and legal standards of proof. Section V concludes.

II. Error Cost Analysis and the Design of Legal Rules

In this Section, we present an error-cost analysis of optimal legal standards of proof. Part A sets out the standard of proof. Part B sets out the loss function used to evaluate standards of proof. Part C derives the optimal statistical standard of proof under specific but, we think, reasonable assumptions about legal standards. Part D examines

¹¹ Hallibuton II, *supra* note 2. See Fisch, Gelbach & Klick, *supra* note 2, for a discussion of the role of statistical significance in that case.

¹² The evaluation of the merits of any statistical analysis requires a normative analysis of the relative frequency and costs of these errors, and should inform the choice of the optimal test and significance level. A full treatment of this last issue is beyond the scope of this paper. In addition, this paper does not address many other issues related to the use and interpretation of statistical models in legal settings, even though there can be interplay between those issues and the way in which statistical inference is performed. For example, at the summary judgment (or trial) phase of a litigation, the issue of liability may be assumed by an expert who is calculating damages with a statistical model. Such an assumption can affect the way in which statistical inference is performed. Another example is that at the certification phase of a class action, where both plaintiffs and defendants generally assume the allegations in the complaint when evaluating the issue of whether impact can be determined with common evidence, a defendant might not put forward its own, competing, statistical model and therefore inferences based on the plaintiffs' model gain more importance than in circumstances where a competing model is available.

mathematical representations of legal standards of proof and the conditions under which these legal standards of proof are optimal standards.

A. The Standard of Proof

We consider the case where the litigants put forward competing hypotheses (H_0 for the defendant and H_1 for the plaintiff) to explain the evidence, which consists of the realization x of a random variable X with a density function $f(X|H_0)$ if hypothesis H_0 is true and $f(X|H_1)$ if hypothesis H_1 is true. Figure 1 illustrates the two conditional sampling distributions under the litigants' competing hypothesis.¹³

FIGURE 1 – The Litigants' Competing Hypotheses and the Standard of Proof x^T

Suppose that the plaintiff has the burden of proof or production. Suppose also that the standard of proof is a threshold value x^T that determines whether the plaintiff has met her burden. Since the sampling distribution of X under the plaintiff's hypothesis H_1 is centered to the left of the sampling distribution of X under the defendant's hypothesis H_0 , realized values of $x \leq (>) x^T$ favor the plaintiff (defendant). Table 1 sets out the decision rule associated with the standard of proof.

¹³ See generally, Louis Kaplow, *Likelihood Ratio Tests and Legal Decision Rules*, 16 AM. L. & ECON. REV. 1 (2014); Louis Kaplow, *Burden of Proof*, 121 YALE L. J. 738 (2012); Louis Kaplow, *On the Optimal Burden of Proof*, 119 J. POL. ECON. 1104 (2011); Edward K. Cheng, *Reconceptualizing the Burden of Proof*, 122 YALE L. J. 1254 (2013); Dominique Demougina & Claude Fluet, *Preponderance of Evidence*, 50 EURO. ECON. REV. 963 (2006).

Table 1: The Standard of Proof

Relationship between Realized Value and Threshold	Legal Inference
$x \leq x^T$	Plaintiff has met her burden of production or proof: Reject H_0 and accept H_1 .
$x > x^T$	Plaintiff has not met her burden of production or proof. Reject H_1 and accept H_0 .

B. Error Cost Analysis and the Loss Function

A preliminary matter is the choice of a “loss function” to evaluate alternative statistical tests and legal standards of proof. The performance of these alternatives will be determined by the specific loss function chosen to evaluate the tests. A common objective used to evaluate evidentiary and other procedural systems is to choose the system that minimizes the sum of two types of costs – expected error costs and the costs of the test.¹⁴ For concreteness, we adopt this approach, while recognizing that reasonable alternatives might exist.¹⁵

The error cost matrix in Table 2 summarizes the outcomes when the standard of proof is x^T , as suggested above. There are two correct outcomes (the shaded boxes on the diagonal) where the true hypothesis is accepted. There are two types of error, Type I error and Type II error. Type I errors (when the plaintiff’s hypothesis H_1 is accepted when H_0 is true) occur at a rate $\alpha(x^T)$ and generate a cost K_I when they occur. The specificity of the test is the rate of correct negatives, and is equal to $1 - \alpha(x^T)$. Type II errors (when the defendant’s hypothesis H_0 is accepted when H_1 is true) occur at a rate $\beta(x^T)$ and generate a cost K_{II} when they occur.¹⁶ The sensitivity of test is the rate of correct positives and is equal to $1 - \beta(x^T)$.

An ideal test would attempt to avoid both types of errors. However, the extent to which an actual test approaches this ideal is limited by available information and the prohibitive costs of acquiring additional relevant and determinative information.

¹⁴ See generally, Richard A. Posner, *ECONOMIC ANALYSIS OF LAW*, 6th Ed. (2003) at 563; See also Ronald J. Allen, *The Error of Expected Loss Minimization*, 2 L. PROB. & RISK 1, 4 (2003).

¹⁵ For alternative approaches to optimal burden of proof rules based on the provision of incentives for underlying behavior, see Kaplow, *supra* note 13; Bruce L. Hay, & Kathryn E. Spier, *Burdens of Proof in Civil Litigation: An Economic Perspective*, 26 J. LEG. STUD. 413 (1997); Chris W. Sanchirico, *The Burden of Proof in Civil Litigation: A Simple Model of Mechanism Design*, 17 INT’L. REV. L. & ECON. 431 (1997).

¹⁶ The loss function does not explicitly include the benefits of correct positive and negative decisions. Following the standard convention in the decision-theoretic literature, the loss parameters are normalized so that gains may be assumed to equal zero.

Table 2 – Error Cost Matrix

	H ₁ is true (P(H ₁))	H ₀ is true (P(H ₀))
Test Positive ($x \leq x^T$)	CORRECT POSITIVE (Sensitivity = $1 - \beta(x^T)$)	FALSE POSITIVE – (Type I Error Rate = $\alpha(x^T)$) (Cost of Type I error = K_I)
Test Negative ($x > x^T$)	FALSE NEGATIVE (Type II Error Rate = $\beta(x^T)$) (Cost of Type II Error = K_{II})	CORRECT NEGATIVE (Specificity = $1 - \alpha(x^T)$)

If $P(H_0)$ and $P(H_1)$ are the prior probabilities that the defendant's and the plaintiff's hypothesis is true respectively, minimization of expected error costs plus the cost of the test is achieved by choosing a standard of proof x^T that minimizes the following loss function:

$$LF(x^T) = P(H_0)[\alpha(x^T)K_I] + P(H_1)[\beta(x^T)K_{II}] + C, \quad (1)$$

where C is the resource cost of applying the standard of proof to the parties and to society. C is assumed not to vary between alternative decision standards, but can be avoided if the legal system uses a standard of proof that does not require evaluation of evidence.

C. The Optimal Standard of Proof

In this section, the standard of proof x^* that minimizes total expected error costs and direct cost is derived. Minimizing the loss function set out in (1) requires taking the derivative of (1) with respect to x^T and setting it equal to zero. This yields the following first order condition:

$$\frac{\partial LF(x^*)}{\partial x^*} = P(H_0) \frac{\partial \alpha(x^*)}{\partial x^*} K_I + P(H_1) \frac{\partial \beta(x^*)}{\partial x^*} K_{II} = 0 \quad (2)$$

Solving the first order condition (3) yields the following condition:

$$-\frac{\frac{\partial \beta(x^*)}{\partial x^*}}{\frac{\partial \alpha(x^*)}{\partial x^*}} = \omega \pi, \quad (3)$$

where $\omega = \frac{K_I}{K_{II}}$ (the ratio of the cost of Type I and Type II errors) and $\pi = \frac{P(H_0)}{P(H_1)}$ (the prior odds in favor of H_0).

Figure 2 adds the error rates $\alpha(x^T) = F(x^T|H_0)$ and $\beta(x^T) = 1 - F(x^T|H_1)$ to Figure 1, where $F(X|H_0)$ and $F(X|H_1)$ are the respective distribution functions associated with the density functions $f(X|H_0)$ and $f(X|H_1)$.

FIGURE 2 – Type I and Type II Error Rates with Standard of Proof x^T

To evaluate the left-hand side of condition (3), note from Figure 2 that the incremental change in the Type I and Type II error rates as the threshold x^T is increased are $\frac{\partial \alpha(x^T)}{\partial x^T} = f(x^T|H_0)$ and $\frac{\partial \beta(x^T)}{\partial x^T} = -f(x^T|H_1)$. Thus, for any choice of the standard of proof threshold x^T (i.e., any choice whether optimal or not), the ratio of first derivatives that enters the optimality condition (3) equals the likelihood ratio evaluated at x^T :

$$-\frac{\frac{\partial \beta(x^T)}{\partial x^T}}{\frac{\partial \alpha(x^T)}{\partial x^T}} = \frac{f(x^T|H_1)}{f(x^T|H_0)} \equiv LR(x^T) \quad (4)$$

Substituting (4) into equation (3) and evaluating at the optimal standard of proof threshold x^* , we see that at the optimal standard of proof, the likelihood ratio must satisfy the following condition:

$$LR(x^*) = \omega\pi, \quad (5)$$

where neither ω nor π varies with the standard of proof. Thus, the optimal standard of proof involves setting the likelihood ratio equal to a constant, whose value depends on the relative costs of Type I and Type II errors (ω) and the prior odds in favor of the null hypothesis (π).

D. Legal Standards and Optimal Standards

In this section, we describe the conditions under which the optimal standard of proof x^* derived in equation (5) coincides with mathematical representations of actual legal standards. Part D.i considers the conditions where the preponderance standard, expressed as posterior odds equal to 1, is an optimal standard that minimizes expected error and direct costs. Part D.ii considers heightened standards of proof.

i. Preponderance

The preponderance standard is a widely-used standard of proof in civil cases. A common mathematical representation of the preponderance standard is that the plaintiff has met his burden under the standard when the probability of the alternative hypothesis is greater than 0.5. For example, Judge Frank Easterbrook, writing for the 7th Circuit, suggested the following interpretation of the preponderance standard:

[T]he trier of fact rules for the plaintiff if it thinks the chance greater than 0.5 that the plaintiff is in the right.¹⁷

This suggests the following absolute probability representation of the preponderance standard:

$$p(H_1|x) > .5. \quad (6)$$

The use of this absolute probability standard as a mathematical representation of the preponderance standard has been criticized by those skeptical of the use of probabilistic models of evidence,¹⁸ and even by those who are not.¹⁹ The latter critics suggest the absolute probability representation be replaced by a ratio that compares the probabilities of the narratives offered by the plaintiff and defendant. In these models, preponderance standard is one that holds for the plaintiff when the ratio of the conditional posterior probabilities evaluated at x is greater than one:²⁰

$$\frac{p(H_1|x)}{p(H_0|x)} > 1 \quad (7)$$

Note that when there are only two competing alternatives, so that $p(H_1|x) + p(H_0|x) = 1$, condition (7) and the absolute probability condition (6), that $p(H_1|x) > .5$, are equivalent statements. In addition, the ratio of the posterior probabilities in (7) equals the conditional odds of H_1 given x .

Using Bayes Rule, condition (7) is equivalent to:

$$\frac{p(H_1|x)}{p(H_0|x)} = \frac{p(x|H_1) P(H_1)}{p(x|H_0) P(H_0)} = LR(x)/\pi > 1 \quad (8)$$

¹⁷ *Brown v. Bowen*, 847 F.2d 342, 345 (7th Cir. 1988).

¹⁸ See, e.g., Ronald J. Allen & Michael S. Pardo, *The Problematic Value of Mathematical Models of Evidence*, 36 J. LEG. STUD. 107 (2007) (discussing reference-class problem); But see Edward K. Cheng, *A Practical Solution to the Reference Class Problem*, 109 COL. L. REV. 2081 (2009) (describing mathematical criteria to address choice of reference class).

¹⁹ Cheng, *supra* note 18.

²⁰ See David H. Kaye, *Statistical Significance and the Burden of Persuasion*, 46 L. & CONTEMP. PROB. 13 (1983).

Solving for $LR(x)$ yields the preponderance rule as a likelihood ratio test:²¹

$$\text{Find for the plaintiff when } LR(x) > LR(x^{PP}) = \pi \quad (9)$$

Comparing the preponderance standard (9) to the optimal standard derived in (5), it is easy to see that the two will coincide when where the cost of Type I and Type II errors are equal, so that $\omega = 1$. Indeed, the Supreme Court, in describing the preponderance standard, has suggested (albeit in dicta) that such a weighting is appropriate in certain civil cases:

In a civil suit between two private parties for money damages, for example, we view it as no more serious in general for there to be an erroneous verdict in the defendant's favor than for there to be an erroneous verdict in the plaintiff's favor. A preponderance of the evidence standard therefore seems peculiarly appropriate for, as explained most sensibly, it simply requires the trier of fact to believe that the existence of a fact is more probable than its nonexistence ...²²

Table 3 depicts the error cost matrix in a civil case that incorporates the assumption that $K_I = K_{II} = K$, so that $\omega = 1$:

Table 3 – Error Cost Matrix in a Civil Case

	H_1 is true ($P(H_1)$)	H_0 is true ($P(H_0)$)
Test Positive ($x \leq x^T$) (liable)	CORRECT POSITIVE (Sensitivity = $1 - \beta(x^T)$)	FALSE POSITIVE – (Type I Error Rate = $\alpha(x^T)$) (Cost of Type I error = $K_I = K$)
Test Negative ($x > x^T$) (not liable)	FALSE NEGATIVE (Type II Error Rate = $\beta(x^T)$) (Cost of Type II Error = $K_{II} = K$)	CORRECT NEGATIVE (Specificity = $1 - \alpha(x^T)$)

Thus, when $\omega = 1$, the preponderance standard (9) is an optimal standard.²³

$$\text{Find for the plaintiff when } LR(x) > LR(x^*) = LR(x^{PP}) = \pi. \quad (10)$$

²¹ See Kaplow, *supra* note 13 at 14.

²² *In re Winship*, 397 U.S. 358 (1970). See also Michael S. Pardo, *The Nature and Purpose of Evidence Theory*, 66 VAND. L. REV. 547, 561 (2013) (citing *Grogan v. Garner*, 498 U.S. 279, 286 (1991) (explaining that the preponderance standard "results in a roughly equal allocation of the risk of error")).

²³ See Cheng, *supra* note 18.

Finally, consider the special case of a diffuse or non-informative prior, i.e., where $P(H_0) = P(H_1) = 1/2$, so $\pi = 1$. Coupled with the assumption that $\omega = 1$, condition (10) becomes:

$$LR(x) > LR(x^*) = LR(x^{PP}) = 1. \quad (11)$$

As Figure 2 is drawn, x^{PP} is the level of proof when the parties' competing densities are equal, reflecting the assumption that $\omega = \pi = 1$. Under these conditions, minimizing the loss function (1) is equivalent to minimizing the total error rate ($\alpha(x^T) + \beta(x^T)$).

For a given choice of proof threshold, using the statistical evidence entails a total loss of $[(K/2)(\alpha(x^T) + \beta(x^T)) + C]$. The best feasible alternative would be to not use statistical evidence and settle for an approach that leads to total error probability of 1, for example, by always determining the issue either for the plaintiff ($\alpha=1, \beta=0, \alpha+\beta=1$) or for the defendant ($\alpha=0, \beta=1, \alpha+\beta=1$); that yields total cost of $K/2$. Thus, total error costs are lower when the test is used whenever:

$$[(K/2)(\alpha(x^T) + \beta(x^T)) + C] < K/2, \quad (12)$$

or, equivalently, whenever

$$\alpha(x^T) + \beta(x^T) < 1 - 2C/K. \quad (13)$$

To illustrate how moving the standard of proof threshold affects the loss function, consider again the evidence and threshold x^T depicted in Figure 2. At x^T , $-\frac{\partial\beta(x^T)}{\partial x^T} =$

$LR(x^T) > 1$, or equivalently $\frac{\partial\alpha(x^T)}{\partial x^T} < -\frac{\partial\beta(x^T)}{\partial x^T}$. Thus, moving the threshold incrementally to the right will increase the Type I error rate by less than the decrease in the Type II error rate, so that the total error cost rate ($\alpha(x^T) + \beta(x^T)$) and thus total error costs $K(\alpha(x^T) + \beta(x^T))$ will fall. This will be true for any threshold $x^T < x^{PP}$. Similarly, for any $x^T > x^{PP}$, moving the threshold incrementally to the left will increase the Type II error by less than the decrease in the Type I error rate. As a result, $x^{PP} = x^*$ when $\pi = \omega = 1$.

We will use this example below to illustrate the differences between legal standards and statistical standards of proof below. The use of the likelihood ratio criterion in (11) is merely an example when there are symmetric error costs and diffuse

²⁴ See Cheng, *supra* note 18.

²⁵ As noted above, this calculation assumes—via the assumption that $\pi = \omega = 1$ —that the legal system is operating with diffuse priors and is indifferent to whether errors operate in favor of plaintiffs or defendants. A system with non-diffuse priors and that was not indifferent would have a boundary condition weighted toward one party type's interests.

prior probabilities, and is not a normative argument that the legal system *should* adopt such priors as a proper presumption, or that such a prior should be adopted in applying relative probability tests.²⁶

ii. Beyond a Reasonable Doubt

The analysis can also be applied to heightened legal standards of proof. Table 4 depicts the error cost matrix in a criminal trial. The criminal error cost matrix differs from the civil matrix in two primary ways. First, commentators have suggested that in the context of a criminal trial, the cost of a Type I error is many times the cost of a Type II error.²⁷ Second, criminal trials incorporate a presumption of innocence. Both the normative weighting of Type I and Type II errors and the presumption of innocence are consistent with the high “beyond a reasonable doubt” (BRD) standard of proof placed on the prosecutor in criminal trials, as well as the robust procedural protections given criminal defendants.²⁸

TABLE 4 – Error Cost Matrix in a Criminal Trial

	H₁ is true (P(H₁))	H₀ is true (P(H₀))
Test Positive ($x \leq x^T$) (guilty)	CORRECT POSITIVE CONVICTING THE GUILTY (Sensitivity = $1 - \beta(x^T)$)	FALSE POSITIVE – CONVICTING THE INNOCENT (Type I Error Rate = $\alpha(x^T)$) (Cost of Type I error = $K_I = K$)
Test Negative ($x > x^T$) (not guilty)	FALSE NEGATIVE LETTING THE GUILTY GO FREE (Type II Error Rate = $\beta(x^T)$) (Cost of Type II Error = K_{II} = ωK , $\omega > 1$)	CORRECT NEGATIVE ACQUITTING THE INNOCENT (Specificity = $1 - \alpha(x^T)$)

Rewriting (5), the optimality condition may be expressed as $LR(x^*)/\pi = \omega$. Recalling that Bayes’s Theorem implies that the posterior odds in favor of the alternative hypothesis, $p(H_1|x)/p(H_0|x)$, equals the ratio $LR(x)/\pi$, the optimal standard of proof induces an equality between the posterior odds and ω . When the costs of Type I errors are presumed to be greater than the cost of a Type II error, $\omega = \frac{K_I}{K_{II}} > 1$. In addition, suppose that presumption of innocence reflects the prior odds π . If we use the Blackstone ratio ($\omega = 10$),²⁹ the posterior odds for the BRD standard of proof to be an

²⁶ See Cheng, *supra* note 18 (arguing for use of a diffuse prior). But see, e.g., Hay & Spier, *supra* note 14 (optimal burden considers the proportion of negligent defendants and the parties’ relative costs of producing evidence).

²⁷ Blackstone wrote that “It is better that ten guilty persons escape than that one innocent suffer.” William Blackstone, COMMENTARIES ON THE LAWS OF ENGLAND (1765).

²⁸ These include *Gideon v. Wainwright*, 372 U.S. 335 (1963) (right to counsel), *Miranda v. Arizona*, 384 U.S. 436 (1966) (notice), *Brady v. Maryland*, 373 U.S. 83 (1963) (requirement that prosecutor disclose material exculpatory evidence in his possession).

²⁹ See note 27, *supra*.

optimal standard $p(H_1|x)/p(H_0|x) = LR(x^{BRD})/\pi = LR(x^*)/\pi = 10$. Thus, to convict, the posterior odds would have to be greater than 10, which equates to requiring that $p(H_1|x) = .909$ to convict when H_1 is the only possible alternative to H_0 . Similarly, if $\omega = 19$, the posterior odds for the BRD standard of proof to be an optimal standard $p(H_1|x)/p(H_0|x) = LR(x^{BRD})/\pi = LR(x^*)/\pi = 19$. This equates to a requiring that $p(H_1|x) = .95$ (a numerical threshold often linked to the “beyond a reasonable doubt” standard) to convict when H_1 is the only possible alternative to H_0 .³⁰

A similar analysis could be applied to model an intermediate civil standard of proof (e.g., the clear and convincing evidence standard used in some civil cases). Note that under this formulation, the optimal critical likelihood ratio $LR(x^*) = \omega\pi$ can differ from the preponderance threshold of $LR(x^{PP}) = 1$ from equation (11) due to asymmetric error costs ($\frac{K_I}{K_{II}} = \omega \neq 1$) or from a non-diffuse prior/presumption ($\frac{P(H_0)}{P(H_1)} = \pi \neq 1$). Thus, an optimal heightened civil standard ($LR(x) > LR(x^{CCE}) = LR(x^*) > 1$) can incorporate the civil presumption ($\pi = 1$) and asymmetric error costs (where $\omega > 1$), symmetric error costs ($\omega = 1$) and a heightened presumption ($\pi > 1$), or both. If such a standard requires that $p(H_1|x) > .75$,³¹ though, it necessarily implies that $\omega = 3$ whenever the null and alternative hypotheses are the only possibilities, so that the clear and convincing standard coincides with the optimal standard when $LR(x)/\pi > LR(x^{CCE})/\pi = LR(x^*)/\pi > 3$. The same analysis can also generate relaxed critical likelihood ratios ($LR(x^{ALT}) = LR(x^*) < 1$) when the asymmetry in error costs and/or the presumption/priors are reversed.

III. Hypothesis Testing and Statistical Inference

A. Null Hypothesis Statistical Testing

Economists, as well as other social scientists have offered expert opinions in litigation that utilize “null hypothesis statistical testing”.³² A null-hypothesis statistical test (“NHST”) is a commonly used method of statistical inference that incorporates

³⁰ See, e.g., Jon O. Newman, *Quantifying the Standard of Proof Beyond a Reasonable Doubt: A Comment on Three Comments*, 5 L. PROB. & RISK 267 (2006) (commenting on quantification of the beyond a reasonable doubt standard as equal to a “95% chance that the defendant is guilty”). See also Joseph L. Gastwirth, *Statistical Reasoning in the Legal Setting*, 46 AM. STATISTICIAN 55, 57 (1992) (Reporting average probabilities associated with legal standards of proof from judicial survey).

³¹ See, e.g., Kaplow, *Burden of Proof*, *supra* note 13 at _.

³² While in decline in some areas of scientific inquiry, NHST using fixed significance levels is still the dominant paradigm in economics and many other disciplines. For a discussion of these issues and the use of alternative statistical inference tools, see, Fidler, et al., *Statistical Reform in Medicine, Psychology and Ecology*, 33 J. OF SOCIO-ECONOMICS 615 (2004); Jonah B. Gelbach, *Type II Error-Sensitive Hypothesis Testing*, University of Pennsylvania Law School Working Paper (2017).

elements of two different statistical approaches pioneered by Ronald A. Fisher³³ and by Neyman & Pearson (“N-P”).³⁴ A NHST analysis used in litigation may offer some statistical result and associated fixed level of statistical significance – usually 5%. The inferential focus of such an analysis is whether the difference between an observed outcome and the outcome under the null hypothesis is statistically different from zero at a 5% level of statistical significance.

NHST, as mentioned, includes elements of statistical methods developed by Fisher and N-P. Fisher’s method was designed as an objective way to evaluate scientific evidence, and focuses exclusively on specifying a “null hypothesis” and examining the data under the assumption that the null hypothesis is true.³⁵ Fisher promoted the use of fixed-significance null-hypothesis testing as a tool for identifying interesting results from scientific research – that is, those results with a low probability of occurring by chance under the null hypothesis. Results that failed to achieve statistical significance were to be ignored.

Fisher’s exclusive focus on the null hypothesis and statistical significance, at the expense of any alternative hypothesis, was intentional. He explicitly rejected the notion that the costs of false negatives (Type II errors) should influence how the threshold of “significance” is set.³⁶

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

Fisher was also the most influential proponent of the use the .05 or 5% significance test, or “rule of two”.³⁷

³³ Ronald A. Fisher, *THE DESIGN OF EXPERIMENTS*, 8th Ed. (1966).

³⁴ Jerzy Neyman & Egon S. Pearson, *On the Problem of the Most Efficient Test of Statistical Hypothesis*, *PHILOSOPHICAL TRANSACTION OF THE ROYAL SOCIETY OF LONDON—SERIES A*, 231 (1933).

³⁵ Fisher, *supra* note 33 at 16 (“In relation to any experiment we may speak of this hypothesis as the ‘null hypothesis’ and it should be noted that the null hypothesis is never proved or established but it is possibly disproved in the course of the experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.”)

³⁶ Fisher, *supra* note 33 at 17. See also Ziliak & McCloskey, *supra* note 3 at 144; Michael I Meyerson & William Meyerson, *Significant Statistics: The Unwitting Policy Making of Mathematically Ignorant Judges*, 37 *PEPP. L. REV.* 771, 823-4 (2010).

³⁷ Ziliak & McCloskey, *supra* note 3 at _ (quoting R. A. Fisher),

The value for which $P=.05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant.

Fisher's 5% significance threshold originally was used as a convenient standard that set out an arbitrary but objective minimum level of precision for declaring research results "important."³⁸ Indeed, under the Fisher approach, arbitrary significance levels other than 5% could be used, did not need to be specified in advance, and could be applied a-posteriori to a given set of data.

Neyman and Pearson's alternative approach to statistical inference attempted to improve on significance testing through a focus on applied decision making and tests of acceptance between competing hypotheses. N-P used the term significance to denote the probability of a false positive or Type I error. Like Fisher, N-P also used "convenient" significance levels, including .05, to set an acceptable level of Type I errors. However, unlike Fisher's approach to significance levels, which could be applied a-posteriori, the N-P approach required an a-priori choice of a significance level. N-P's approach also required the specification of an alternative hypothesis in addition to specifying the null.

N-P's explicit consideration of alternative hypotheses in turn allowed consideration of "effect size" as well as the explicit consideration of the rate of Type I and Type II errors and the ability to establish tests that would have the highest power. Given the choice of the rate of Type I error and an alternative hypothesis, Type II errors were to be controlled by ensuring the data was of sufficient sample size given the effect size.

Descriptions and definitions of NHST vary, but can be described as a general set of "cookbook" procedures used to carry out statistical testing rather than a precise methodology for conducting statistical inference. While application of NHST varies, it does encompass a certain set of procedures that incorporate elements of both Fisher's and N-P's approaches without attempting the difficult task of reconciling inconsistent approaches into a unified methodology.³⁹

NHST generally begins with the specification of a null hypothesis and a choice of an arbitrary significance level, with the 5% significance level commonly used. An alternative hypothesis is also specified. However, the level of specificity of the alternative hypothesis varies. The most common or what we will refer to as the "standard" approach is to use a non-specific alternative hypothesis (e.g., when the alternative is the negation of the null hypothesis or a one-directional alternative, known as a composite alternative). If the alternative hypothesis is not specific, then NHST will more closely resemble Fisher's approach to significance testing.

³⁸ Id at 45-6. See also Meyerson & Meyerson, *supra* note 36.

³⁹ For a discussion of the two approaches being complementary rather than contradictory, see E. L. Lehmann, *The Fisher, Neyman-Person Theories of Testing Hypotheses: One Theory or Two?*, 88 JASA 424 (1993).

A more specific alternative hypothesis will allow consideration of statistical power and effect size, and will more closely resemble the N-P approach to tests of acceptance. As will be set out below in more detail, our approach to applying NHST to make legal inferences will more closely follow a modified N-P approach using tests of acceptance. Our approach relies on specifying a specific alternative hypothesis and allowing for consideration of statistical power and effect size. However, because the sample size is effectively fixed in many legal settings, our approach will also emphasize implementation of the tradeoff between Type I and Type II errors implied by the applicable legal standard for a given sample size.

B. Null and Alternative Hypotheses

In order to illustrate both the NHST approach to hypothesis testing and to compare this approach to the general error cost framework set out in Section II, we use the following example throughout the remainder of the paper.⁴⁰ Suppose that an employer fills N vacancies by hiring thirty percent women and seventy percent men.⁴¹ A non-hired female applicant sues the employer under Title VII for discriminating against women.⁴² The complaint alleges that the employer's hiring practices had a disparate impact on women.⁴³ The plaintiff's initial burden of production is to prove, by a preponderance of evidence, a *prima facie* case of statistical discrimination.⁴⁴ In such cases, this is achieved by showing that the number of women hired is lower than expected, and that the difference is significant given the court's choice of significance level.⁴⁵

⁴⁰ The basic structure of the example is taken from Daniel L. Rubinfeld, *Econometrics in the Courtroom*, 85 COL. L. REV. 1048, 1058 (1985).

⁴¹ This paper does not address the EEOC 4/5 rule for disparate impact, which examines the effects of employment selection criteria and finds a *prima facie* case if the ratio of selection rates of women to men is less than .8. For a comparison of the EEOC 4/5 rule and statistical significance, see Peresie, *supra* note 6, and Finkelstein & Levin, *supra* note 6 at ___. The use of the binomial probability models used in the example in the paper may be relevant if the selection mechanism responsible for the disparate treatment is not observable. For example, suppose that a firm fills ten slots by offering the jobs to the first ten applications pulled out of a large drum. However, the employee who is responsible for placing the applications in the drum systematically places the applications of some women at the bottom of the drum. In this case, the observed selection rate for men and women will both be 1. For an example of such a process in the context of race and jury selection, see *Swain v. Alabama*, 380 U.S. 202 (1965) (black persons in the jury pool had different colored tickets than whites).

⁴² Title VII of the Federal Civil Rights Act of 1964, 42 U.S.C. §§ 2000e et seq. See generally, Peresie, *supra* note 6; Garaud, *supra* note 6.

⁴³ See *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971).

⁴⁴ *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642 (1989).

⁴⁵ See Garaud, *supra* note 6 at 467.

The inferential problem for both legal and statistical analysis is to decide when the observed outcome (a lower than expected number of women hired) favors an alternative hypothesis that the employer is discriminating against women over the null hypothesis that the employer hired employees without regard to the applicant's sex.⁴⁶ Under a fixed-significance level NHST approach to statistical inference, the null hypothesis of no discrimination is rejected in favor of an alternative hypothesis of discrimination when the observed outcome x is less than a threshold x^C determined by an arbitrarily set fixed level of statistical significance. As set out in Section II, evidentiary standards and burdens of proof and production, such as the preponderance of evidence, also use threshold levels of the same observed outcome to determine liability. A key issue is to analyze how these alternative threshold values are determined, and to provide a framework for evaluating both the consistency and performance of alternative thresholds.

As noted above, the NHST approach to statistical inference compares one hypothesis (the "null" hypothesis denoted H_0) to an "alternative" hypothesis (denoted H_1) or to a set of alternative hypotheses. We first describe in more detail the concepts of null and alternative hypotheses, and present an example that will be used to illustrate the concepts and alternative approaches to statistical inference.

i. The Null Hypothesis (H_0)

The null hypothesis is the hypothesis to be tested; in the parlance of statistics, the null hypothesis will either be rejected or fail to be rejected. It is commonly defined as the hypothesis that any observed difference (for example, between two sample means or a sample mean and baseline value) is due to random chance and not systematic causes.

In the gender discrimination example, the null hypothesis (H_0) is that the employer hires N employees without regard to sex from a large pool of equally qualified applicants. In the pool of qualified applicants, p_0 is the proportion of qualified applicants that are female, and $1 - p_0$ is the proportion of qualified applicants that are male. In terms of the defendant's hypothesis discussed in Section II, the null and defendant's hypothesis would be the same.

The firm actually hires x females and $N - x$ males. Let X be the random variable representing the number of women hired, so that under the null hypothesis, X has

⁴⁶ The inquiry here is a preliminary one – whether the plaintiff has met his initial burden of production for a *prima facie* case. The general principles analyzed in this paper can be applied to other types of preliminary thresholds such as the plausibility standard to decide motions to dismiss in *Bell Atlantic Corp. v. Twombly*, 550 U.S. 544 (2007) and *Tellabs Inc. v. Makor Issues & Rights*, 551 U.S. 308 (2007), the standard for summary judgment under *Matsushita Electric Industrial Co., Ltd. v. Zenith Radio Corp.*, 475 U.S. 574 (1986), or the burden of proof for final adjudication. This issue is discussed in Section IV.E *infra*.

approximately a binomial sampling distribution, with a probability density function given by:⁴⁷

$$f(x; N, p_0) = \binom{N}{x} p_0^x (1 - p_0)^{N-x}. \quad (14)$$

The binomial density function gives the probability of hiring exactly x females for the N open slots. The expected number of females hired if H_0 is true is $\mu_0 = Np_0$, and the variance equals $\sigma_0^2 = Np_0(1-p_0)$. For example, if the firm hires $N = 10$ qualified workers without regard to the sex of the applicant from a qualified pool that is half female and half male, the probability that a given hire will be a female is $p_0 = .5$ under H_0 .⁴⁸ Thus, under the assumption that the relevant pool of qualified workers is half male and half female, the null hypothesis is equivalent to a null hypothesis that $p_0 = .5$. From the well-known properties of the binomial distribution, the expected number of females hired equals $\mu_0 = Np_0 = 5$, and the variance equals $\sigma_0^2 = Np_0(1-p_0) = 2.5$. Figure 3 depicts the sampling distribution of the number of women hired under H_0 , i.e., when $p_0 = .5$ and $N = 10$. In the figure, the horizontal axis measures the number of women that could be hired, x , which can range from 0 to 10. The vertical axis measures the probability X takes on each of these values, when the null hypothesis is true.

The null sampling distribution in Figure 3 illustrates that even when the firm hires applicants without considering the sex of the applicant, it is not the case that a given hiring cycle always will produce an observed outcome where half of new hires are men and half are women. In fact, when an employer selects qualified workers without regard to sex, and the qualified pool is half men and half women, we would expect that outcome to occur a shade less than one quarter of the time. Examining Figure 1, the probability that exactly five women and five men would be hired when H_0 is true ($p(X=5|H_0)$) equals 0.2461 (or 24.61%).

In the hypothetical example set out above the employer hired 70% men and 30% women. With $N = 10$, this means 3 women and 7 men would be hired for the ten

⁴⁷ For a discrete random variable, such as a count of the number hired, the usual term is the probability mass function; the term probability density function is used when the random variable of interest has a continuous distribution. Since nothing important will be changed by ignoring this distinction for purposes of our discussion, we use the term “density” throughout. Note also that the sampling distribution for the number of count variable (the number of women hired, X) is only approximately binomially distributed because the pool of qualified applicants is sampled without replacement. The precise distribution is a hypergeometric one, but using a binomial distribution here will simplify the discussion without changing anything important (A general rule of thumb is that the binomial sampling distribution for counts can be used when the population is at least 20 times as large as the sample, a condition likely met in most applications).

⁴⁸ Our examples set $p_0 = .5$. When the qualified pools are not 50% female and 50% male, then p_0 will deviate from .5. See *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, __ (1989) (citing *Hazelwood School District v. United States*, 433 U.S. 299, 308 (1977)), “[t]he proper comparison [was] between the racial composition of [the at-issue jobs] and the racial composition of the qualified . . . population in the relevant labor market.”

vacancies. The probability of this outcome ($x = 3$) under the null hypothesis is 11.72%, and the probability that three or fewer women will be hired equals 17.18% (the sum of the four relevant probabilities). The “significance” of these numbers will be addressed below.

FIGURE 3 – Binomial Sampling Distribution of the Number of Women Hired (X) Under the Null Hypothesis ($p_0 = .5$ and $N = 10$)

ii. The Alternative Hypothesis (H_1)

Under NHST, the alternative hypothesis is the hypothesis that is favored when the null is rejected. There are several types of alternative hypotheses. These include non-specific compound hypotheses that are used in the “standard” approach to NHST, which include non-directional alternative hypotheses, one-and two-tailed directional alternative hypotheses, and point hypotheses.⁴⁹ Non-directional alternative hypothesis, such as the negation of the null hypothesis, are often used in settings where the only concern is whether the null hypothesis is true. Directional alternative hypotheses can be one-tailed, where only one tail of the region of rejection is of concern, or two tailed, where both regions of rejections are of concern. In the discrimination example set out above, a non-directional alternative hypothesis would be that $p_0 \neq .5$. But such a two-tailed alternative hypothesis makes no sense in an employment discrimination lawsuit. To see this,

⁴⁹ As noted above, hypotheses that encompass more than one possibility are called composite hypotheses. Composite hypotheses may include any number of point hypotheses.

suppose the employer-defendant actually hired 9 women out of 10 total hires. Even though this is a very unlikely outcome if the employer hires in a sex-neutral way, it is unlikely in a way that would provide evidence the employer *favors* female plaintiffs, rather than discriminating against them. No court would direct or uphold a verdict against the employer under such conditions. The relevant alternative hypothesis would be a directional one, often known as a one-tailed alternative hypothesis (that women are hired less than half the time, or $p_1 < .5$).

Our analysis in this paper will rely on the third general type of alternative hypothesis, point or specific alternative hypotheses.⁵⁰ Point alternative hypotheses are hypotheses where the sampling distribution under the alternative hypothesis is a fully defined distribution with no unknown parameters. While their explicit use in empirical analysis is relatively rare, specific alternative hypotheses are fundamental to the N-P approach to statistical inference, and the associated determining concepts of statistical power, effect size, and optimal likelihood ratio tests under the Neyman-Pearson lemma.⁵¹ Use of point or specific alternative hypotheses also is helpful to understanding the relationship between statistical inference and legal inference discussed in Section II. Both Kaplow's approach to legal decision rules—based on Neyman-Pearson likelihood ratio tests⁵²—and Cheng's relative probability approach use specific alternative hypotheses.⁵³ And recent approaches to legal inference and burdens of proof incorporate specific explanations of the evidence, the legal analog of point hypotheses. Pardo & Allen's explanation-based reasoning and the process of inference to the best explanation make use of a finite set of specific explanations, generated by the parties in the first stage of the two-stage explanation-based inferential process.⁵⁴

Table 5 provides example data for a null hypothesis and three specific alternative hypotheses based on the discrimination example. The Table lists the values of the

⁵⁰ This paper does not consider composite hypotheses. For an approach to balancing Type I and Type II error costs in such situations, see Gelbach, *supra* note 32.

⁵¹ In terms of the example set out above, the Neyman-Pearson lemma states that when performing a test between two specific hypotheses, H_0 and H_1 , the likelihood ratio test which rejects H_0 in favor of H_1 when the likelihood ratio evaluated at x is greater than some threshold likelihood ratio is the most powerful test for a given significance level. Under some conditions, the Neyman-Pearson lemma can sometimes be generalized to composite alternative hypotheses (though it generally cannot when the alternative is simply the two-tailed negation of a point-null hypothesis).

⁵² See Kaplow, *supra* note 13.

⁵³ Cheng, *supra* note 18.

⁵⁴ Michael S. Pardo & Ronald J. Allen, *Juridical Proof and the Best Explanation*, 27 L. & PHIL. 223, _ (2008) (discussing factors that determine the inferential interests and the level of detail of the explanations); Ronald J. Allen & Alex Stein, Evidence, *Probability, and the Burden of Proof*, 55 ARIZ. L. REV. 557 (2013) (distinguishing “relative plausibility theory” from mathematical approaches to evidence); Pardo, *supra* note _ (explaining how inference to the best explanation explains evidentiary law).

sampling distribution density function for the null and each of the alternative distributions, as well as the ratio of the density function for each of the specific alternatives relative to the null, the likelihood ratio $LR(X)$.⁵⁵

TABLE 5: Null and Specific Alternative Hypotheses

Hypothesis	a. Null (H_0)	b. Alt. (H_1^b)		c. Alt. (H_1^c)		d. Alt. (H_1^d)	
Probability female is hired	$p_0 = .5$	$p_1^b = .444$		$p_1^c = .375$		$p_1^d = 0$	
Number of Women Hired (X)	$f(X H_0)$	$f(X H_1^b)$	$LR(X)$	$f(X H_1^c)$	$LR(X)$	$f(X H_1^d)$	$LR(X)$
0	0.0010	0.003	3.0	0.009	9.0	1	1000
1	0.0097	0.023	2.3	0.055	5.5	0	0
2	0.0439	0.081	1.8	0.147	3.3	0	0
3	0.1172	0.173	1.5	0.236	2.0	0	0
4	0.2051	0.241	1.2	0.248	1.2	0	0
5	0.2461	0.231	0.9	0.178	0.7	0	0
6	0.2051	0.154	0.7	0.089	0.4	0	0
7	0.1172	0.070	0.6	0.031	0.3	0	0
8	0.0439	0.021	0.5	0.007	0.2	0	0
9	0.0097	0.004	0.4	0.001	0.1	0	0
10	0.0010	0.0003	0.3	5.5E-05	0.1	0	0

The first column of Table 5 provides probabilities under the null hypothesis, described above, the hypothesis that the firm hires qualified employees without regard to gender. The last column (Column d) represents the specific alternative hypothesis that the firm does not hire women (H_1^d). The number of women hired is generated by a binomial sampling distribution with parameter $p_1^b = 0$. Column c represents the specific alternative hypothesis that the firm discards applications submitted by married women, and thus when it hires women, hires only those who are single (H_1^c). Assuming 40% of qualified females are married and 60% are single, the probability that a woman is hired under this alternative hypothesis is 3/8 (or .375).⁵⁶ The number of women hired will be

⁵⁵ See note 51, *supra*. Rubinfeld, *supra* note 40 at _.

⁵⁶ Under H_1^c , suppose that all applications from qualified married women are discarded. The remaining set of applications is then forwarded to the hiring committee, and 10 are randomly picked. Maintaining the assumption that 50% of the applicants are women and 50% are men, under H_1^c , 40% of the women (2/5) and 20% of the total files (2/10) are discarded, and $p_1^c = .3/.8 = .3750$. The same probability can be derived from a sequential process where applications from qualified married women that are randomly picked are not considered and thrown back into the large pile of applications, while applications from qualified men and single women are approved.

generated by a binomial sampling distribution with parameter $p_I^c = .375$. Finally, Column (b) represents the specific alternative hypothesis that the firm only hires women without children (H_1^b). Assuming 20% of qualified females have children, the probability that a woman will be hired under this alternative hypothesis is 4/9, or $p_I^b = .444$.⁵⁷ Figure 4 shows the binomial sampling distributions for the four hypotheses in Table 1.

FIGURE 4 – The Null and Specific Alternative Distributions

C. Null Hypothesis Statistical Testing and Fixed Significance Levels

When NHST is implemented with a fixed significance level, the null hypothesis is rejected when the observed number of females hired has a p-value less than α (the fixed significance level). The p-value is the probability of observing an outcome equal to or more extreme than the observed value when the null hypothesis is true.⁵⁸ Consider the discrimination example set out above. Examining Figure 3, the one-tailed p-value associated with observing 3 or fewer women hired when the null hypothesis is true equals .1718. If the conventional $\alpha = .05$ significance level is applied to this outcome, the deviation of 3/10 females hired from the expected 5/10 under the null hypothesis is not statistically significant (as the p-value = .172 > $\alpha = .05$). In fact, the deviation of 2/10 females hired from the expected 5/10 is not statistically significant at the conventional .05 level either (with a p-value = .055 > $\alpha = .05$). Only observed outcomes where zero or one female out of ten would be statistically significant based on a test that required p-values below $\alpha = .05$.

⁵⁷ Under H_1^b , suppose that all applications from women with children are discarded. Thus, under H_1^b , 20 percent (1/5) of the women and 10 percent of the total files (1/10) are screened out, and $p_I^b = 4/9 = .444$.

⁵⁸ See Federal Judicial Center, REFERENCE MANUAL ON SCIENTIFIC EVIDENCE, 3rd ed. (2011) at 250.

FIGURE 5 – Threshold for Significance Level $\alpha = .05$

Thus, the threshold of statistical significance based on p-values $< .05$ is between one and two females hired out of 10, and is illustrated in Figure 5. Figure 5 shows the distribution of outcomes when the null is true. When the null hypothesis is true, the $\alpha = .05$ significance test erroneously rejects the null when 0 or 1 females out of 10 are hired. As Table 3 indicates, these outcomes are false positives or Type I errors. This type of error occurs with probability .0107.

Setting the significance level at α ensures that the Type I error rate will be no greater than α . In the discrete binomial example illustrated in Figure 5, setting the significance level $\alpha = .05$ results in a Type I error rate of less than .05. This is due to the discrete nature of the binomial distribution. Figure 6 further illustrates how setting $\alpha = .05$ fixes the Type I error rate to be equal to $\alpha = .05$. The top panel of Figure 6 superimposes the normal approximation onto the binomial distribution in the case where $N = 10$.⁵⁹ With a normal distribution, an $\alpha = .05$ significance level corresponds to a one-

⁵⁹ The normal approximation with $N = 10$ is used for convenience to illustrate why α is the rate of Type I error. However, in general, the normal approximation is not accurate for small values of N . The normal approximation is appropriate when $Np > 10$ and $Np(1-p) > 10$, conditions which are not met in this example.

tailed critical value equal to 2.40 (or 1.645 standard deviations below the mean).⁶⁰ The bottom panel of Figure 6 shows the left one-tailed normal distribution cutoff for a $\alpha = .05$ significance level. When the null hypothesis is true, rejecting the null hypothesis when $x < x^C = 2.401$ will erroneously reject the null hypothesis 5% of the time when x is a normally distributed continuous variable.⁶¹

FIGURE 6 – The Normal Approximation of the Null Distribution and $\alpha = .05$ Normal Distribution Critical Values for a One-Tailed Test

D. Type II Error and Statistical Power with a Fixed- $\alpha = .05$ NHST

The fixed- $\alpha = .05$ NHST considers only Type I error and the sampling distribution under the null hypothesis. Thus, in a fixed N setting, this approach ignores the rate and costs of Type II error as well as the distribution of the alternative hypothesis. Determining the rate of Type II errors associated with a fixed- $\alpha = .05$ NHST requires

⁶⁰ Using the standard deviation from the underlying binomial distribution, with $N = 10$ and $p_0 = .5$, which equals $[Np_0(1-p_0)]^{1/2} = [10(.5)(.5)]^{1/2} = .158$, 1.645 standard deviations equals 2.599, and 1.96 standard deviations equals 3.097. The left-hand cutoff for a $\alpha = .05$ one tailed test equals $5 - 2.599 = 2.401$ and the left-hand cutoff for a two-tailed test equals $5 - 3.097 = 1.903$.

⁶¹ When the null distribution can be approximated by a normal distribution with known parameters ($\mu_0 = Np_0$, $\sigma_0^2 = Np_0(1-p_0)$), a standardized test statistic ($Z = (x - \mu_0)/\sigma_0$) can be calculated and compared to the applicable cutoff for the fixed α null hypothesis test. For an $\alpha = .05$ one tailed test, the null hypothesis is rejected if $|Z| > 1.645$.

explicit consideration of the applicable sampling distribution under the relevant alternative hypothesis. To illustrate the power associated with a fixed- $\alpha = .05$ NHST, suppose that the plaintiff alleges that the firm's hiring practices discriminate against married women. The null hypothesis remains the same and the alternative hypothesis is now described by H_1^c . Under H_1^c , the plaintiff alleges that the firm discriminates against married women, who make up 40% of the qualified pool of females. The relevant sampling distribution is a binomial distribution with parameter $p_1^c = .375$.

FIGURE 7 – Distribution of the Number of Women Hired under the Null and H_1^c , Fixed- $\alpha = .05$ NHST

Figure 7 shows both the null and the alternative distribution of X given H_1^c when the firm attempts to fill ten vacancies ($N = 10$). As shown above, a fixed $\alpha = .05$ NHST will result in a critical value that rejects the null hypothesis of no discrimination if one or zero women are hired. This results in a probability of erroneously rejecting the null when the null is true equal to .0107. This Type I error rate is calculated by adding the probabilities of observing less than 2 women hired when the null is true, or adding the frequencies from the null distribution to the left of the cutoff. A Type II error occurs when the fixed- $\alpha = .05$ NHST fails to reject the null when the alternative is true. The Type II error rate is calculated by adding the probability of observing 2 or more women hired when the alternative is true, or adding the frequencies from the alternative distribution to the right of the cutoff. The Type II error rate of the fixed- $\alpha = .05$ NHST is .9363 in this case, so under H_1^c the sum of Type I and Type II error rates is 0.947.

IV. Legal and Statistical Significance Compared

In this section, the error cost analysis discussed in Section II is compared to fixed-significance level NHST discussed in Section III. The fixed-significance level NHST is conceptually different than the error cost analysis. The standard approach to fixed- $\alpha = .05$ NHST begins with the application of the “standard” or “conventional” significance level of 5%. The fixed- $\alpha = .05$ NHST considers only Type I error rates and the null distribution in fixing the Type I error rate at 5%. The rate and costs of Type II errors as well as the distribution of the alternative hypothesis are ignored.

In terms of the error cost matrix illustrated in Table 2, fixing the significance level at α is equivalent to setting the Type I error rate at α irrespective of the level of Type II error β . Table 6 illustrates the error cost matrix for the disparate impact discrimination example when a fixed $\alpha = .05$ NHST is used. Suppose that the legal standard to prove a *prima facie* case of disparate impact is a statistically significant deviation in the number of women hired relative to the expected number under the null hypothesis. A Type I error is the rejection of the null hypothesis (a positive test, finding there is a significant deviation) when the null hypothesis is true (i.e., the firm hires qualified workers without regard to sex). A Type II error is the failure to reject the null hypothesis (a negative test, failing to reject the null) when the null hypothesis is false (the firm discriminated against women when making employment decisions).⁶² The thresholds for significance levels are set without consideration of the rate of Type II error or any reference to or consideration of the potential alternative distributions.⁶³

TABLE 6 – Error Cost Matrix in a Fixed- α NHST

	Alternative is true (Firm discriminates against women)	Null is true (Firm picks employees without regard to sex)
Test Positive ($x \leq x^c$) (statistically significant)	CORRECT POSITIVE CORRECT INFERENCE OF STATISTICAL DISCRIMINATION (Power/Sensitivity = $1 - \beta(x^c) = ?$)	FALSE POSITIVE – INCORRECT INFERENCE OF STATISTICAL DISCRIMINATION (Type I Error Rate = $\alpha(x^c) = .05$)
Test Negative ($x > x^c$) (not statistically significant)	FALSE NEGATIVE ERRONEOUS NO INFERENCE OF STATISTICAL DISCRIMINATION (Type II Error Rate = $\beta(x^c) = ?$)	CORRECT NEGATIVE CORRECT NO INFERENCE OF STATISTICAL DISCRIMINATION (Specificity = $1 - \alpha(x^c) = .95$)

In contrast, the error cost analysis presented in Section II considers both Type I and Type II error rates, the distributions of both the null and alternative hypotheses, as well as the costs associated with avoiding the two types of errors. As will be discussed below, given these differences, the two approaches generally will produce different tests and different outcomes. To the extent that legal standards and procedures consider both the costs of Type I and Type II errors, it would be mere coincidence if a standard that fixes the rate of one type of error (Type I errors), without considering either the relative cost or the rate of Type II errors, just happened to minimize the applicable loss function.

For expositional purposes, we focus in this section on the civil standard of preponderance of the evidence with diffuse priors. Under these conditions, the

⁶² There is a preference for using the term “do not reject” or “fail to reject” rather than “accept” the null hypothesis. See, e.g., Robert V. Hogg & Elliot Tanis, *PROBABILITY AND STATISTICAL INFERENCE*, _ ed. at 243.

⁶³ See Ziliak & McCloskey, *supra* note 3.

preponderance test is identical to the $LR(x^*) = 1$ optimal test that minimizes sum of the error rates $\alpha + \beta$. This section compares this mathematical formulation of the civil legal standard under these conditions and the fixed- α statistical test.

A. Fixed- $\alpha = .05$ NHST Versus $LR(x^*) = 1$, Minimum $\alpha + \beta$ Tests

Figure 8 depicts the properties of the $LR(x^*) = 1$ test that minimizes $\alpha + \beta$ when the null hypothesis of no discrimination ($p_0 = .5$) is tested against the alternative H_1^c hypotheses that the firm does not hire married women ($p_1^c = .375$) and $N = 10$. The $LR(x^*) = 1$ test rejects the null hypothesis when less than five women are hired. In contrast, as shown in Figure 7, the Fixed- $\alpha = .05$ NHST test reject the null hypothesis when less than two women are hired. Thus $x^C < x^{PP} = x^*$.

Note that while the $LR(x^*) = 1$ test has a higher Type I error rate ($\alpha(x^*) = .3770$) than the fixed- $\alpha = .05$ NHST depicted in Figure 5 ($\alpha(x^C) = .0107$), it will have much lower Type II error rate ($\beta(x^*) = .3057$) than the fixed- $\alpha = .05$ NHST (with a $\beta(x^C) = .9363$). As a result, the sum of $\alpha(x^*) + \beta(x^*) = .6827$ for the $LR(x^*) = 1$ test is lower than the sum of $\alpha(x^C) + \beta(x^C) = .947$ for the fixed- $\alpha = .05$ NHST.⁶⁴

FIGURE 8 – Distribution of the Number of Women Hired under the Null and H_1^c , $LR(x^*) = 1$ Minimum $\alpha + \beta$ Test

The same tradeoffs are illustrated in Figure 9. On the horizontal axis is the specificity of the test, or the rate at which the test correctly fails to reject the null hypothesis when the null is true. The vertical axis measures the power of the test, or the

⁶⁴ In addition to producing higher error costs, the $\alpha = .05$ fixed-significance level test would not satisfy the boundary condition (3') for a cost-effective test, and its use would result in losses plus direct costs that are higher than if the test were simply not performed. This is due to the very high Type II error rate associated with this test.

rate at which the test correctly rejects the null when the alternative hypothesis is true. A perfect test would have power and specificity of 1, and is located at the upper right hand corner of the Figure. Under the circumstances posited in the example, such tests are unattainable with a finite number of observations. Given the cost and limitations of information in determining an outcome, there will be a set of feasible tests which differ in terms of the rates of errors and the costs associated with administering and participating in the test. Determining which of the many feasible, but imperfect, tests is preferred will depend on the relative cost of Type I and Type II errors.

FIGURE 9 – Loss Function and Feasible Significance Tests for H_1^c

Combinations of power and specificity that are attainable with a finite ($N = 10$) number of observations are depicted in the Figure, and lie along the curve labeled the “production possibilities curve.” This curve contains the possible combinations of specificity and power as the threshold cutoff of the number of females hired is moved from greater than 10 (unconditionally accept the null) to less than zero (unconditionally reject the null). The former lies at the bottom right corner of the Figure, and yields a test with perfect specificity and zero power. The latter lies at the top left corner of the Figure, and yields a test with zero specificity and perfect power. These tests are in the set of non-informative or “useless” tests, *i.e.*, those that have a likelihood ratio equal to one for all values of X . Such tests return results (positive or negative) that are the same whether the null or alternative is true.⁶⁵

⁶⁵ Such tests include per se rules (tests that always come out one way), as well as random tests (e.g., deciding guilt or innocence through the outcome of a coin toss). See Finkelstein & Levin, *supra* note 6 at _.

The Figure depicts, in the grey shaded area labeled “cost effective tests”, the set of tests that satisfy the boundary condition (13) under the assumption that $2C/K = .3$. Finally, Figure 9 depicts the test on the production possibilities curve that minimizes $\alpha + \beta$. Under the assumption that $\pi = \omega = 1$, tests with equal levels of expected losses given by equation (1) lie along line with slope of -1. Among the feasible tests, the test that minimizes the total error cost rate $\alpha + \beta$ is the test where the slope of the production possibilities curve is tangent to the minimum $\alpha + \beta$ line (the dashed line in the Figure). This test is the $LR(x^*) = 1$ test that rejects the null hypothesis when the number of females hired x is less than 5. The $LR(x^*) = 1$ test, based on the binomial probabilities from the null and alternative H_1^c binomial distributions, yields a Type I error rate $\alpha(x^*) = .377$, a Type II error rate $\beta(x^*) = .306$, and a total error rate $\alpha(x^*) + \beta(x^*) = .683$. Under the assumption that $2C/K = .3$, this test satisfies the boundary condition (13), as $\alpha(x^*) + \beta(x^*) = .683 < 1 - .3 = .7$.

Figure 10 shows where the $LR(x^*) = 1$ test and the fixed- $\alpha = .05$ NHST lie on the production possibilities curve under H_1^c (where $p_1^c = .375$) when $N = 10$. The fixed- $\alpha = .05$ NHST is located on the production possibilities curve in the dark shaded region closest to the dashed horizontal line labeled $\alpha = .05$. As discussed above, this rule rejects the null hypothesis when less than two women are hired and generates a sum of Type I and Type II error rates ($\alpha(x^c) + \beta(x^c)$) close to 1, the measure of a non-informative or useless test. In contrast, the $LR(x^*) = 1$ test on the production possibilities curve that minimizes the sum of the error rates ($\alpha(x^*) + \beta(x^*)$) produces a lower total error cost rate equal to .6827, and lies on an $(\alpha + \beta)$ -line with a lower total error rate. The test also lies within the cost-effective range.

FIGURE 10 – $LR(x^*) = 1$, Minimum $\alpha + \beta$ Test versus Fixed- $\alpha = .05$ NHST

B. α , β and the Relevant Alternative Hypothesis

The properties of the $LR(x^*) = 1$ test depend on the specified alternative hypothesis, and can generate levels of α that are less than or greater than .05. To illustrate this point, consider Alternative Hypothesis d (H_1^d), the hypothesis that the firm does not hire women, so that $p_1^d = 0$. Figure 11 shows distributions of the number of women hired for both the null and Alternative Hypothesis (H_1^d) on the same graph. A fixed- $\alpha = .05$ NHST (which rejects the null when fewer than 2 women are hired) produces a Type I error rate of .0107. Because the distribution under Alternative Hypothesis (H_1^d) has 100% its mass at 0, the probability of 2 or more women being hired under H_1^d is zero. Thus, this test produces a Type II error rate of zero. While the fixed- $\alpha = .05$ NHST produces low Type I and zero Type II errors, a superior test exists which lowers Type I error without producing Type II error. This is the test that rejects the null hypothesis when zero women are hired. Compared to the fixed- α test, this test does not increase β , but it does reduce α by .01.

FIGURE 11 – Distribution of the Number of Women Hired under the Null and Alternative d

Figure 12 illustrates the $LR(x^*) = 1$ test that minimizes $\alpha + \beta$ for all the three alternative hypotheses contained in Table 1. In addition to the production possibility curve for Alternative Hypothesis H_1^c depicted in Figure 9, Figure 12 also contains the production possibility curve for the other two alternative hypotheses listed in Table 1, H_1^b (the firm does not hire women with children) and H_1^d (the firm does not hire women). The $LR(x^*) = 1$ test that minimizes $\alpha + \beta$ for H_1^d has high power and high specificity, and is a near perfect test that lies in the upper right hand corner of the Figure. The $LR(x^*) = 1$ test that minimizes $\alpha + \beta$ for H_1^b , in contrast, is not cost effective. For example, the $LR(x^*) = 1$ test that minimizes $\alpha + \beta$ under H_1^b has the same $\alpha(x^*) = .377$ as the test that minimizes $\alpha + \beta$ under H_1^c . However, the test has a lower power, and thus a higher β . This is because the test under H_1^b is attempting to discern between two hypotheses that are much closer to each other ($p_0 = .5$ versus $p_1^b = .444$, a difference of .056) that the test

under H_1^c ($p_0 = .5$ versus $p_1^b = .375$, a difference of .125). Thus, for tests of equal significance, H_1^b will have lower power than H_1^c .

FIGURE 12 – Alternative Hypotheses and the Nature of the $\alpha + \beta$ Minimizing Test

The $LR(x^*) = 1$ tests under H_1^b , H_1^c , and H_1^d produce Type I error rates of $\alpha(x^*) = .38$, $.62$, and less than $.001$ respectively.

C. α , β and N

The rate of the errors α and β are affected by the sample size, N , which in turn determines the standard deviation of the underlying distributions of the null and alternative hypotheses. For a fixed- α NHST test, a higher N will result in greater power (a lower β), but—by design—no change in α . Figure 13 illustrates the production possibilities curve of the $LR(x^*) = 1$ test for H_1^c for different levels of N . In addition to the production possibilities curve for $N = 10$ that is illustrated above, Figure 13 provides the production possibilities curve for $N = 100$ and $N = 1000$. As shown above, the fixed- $\alpha = .05$ NHST for H_1^c and $N = 10$ yields error rates $\alpha(x^C) = .0107$ and $\beta(x^C) = .9363$. When $N = 100$, $\alpha(x^C) = .04$ and $\beta(x^C) = .204$ and when $N = 1000$, $\alpha(x^C) = .047$ and $\beta(x^C) = .0001$.

For the $LR(x^*) = 1$ tests, the minimum α and β decrease as N increases. While the minimum $\alpha + \beta$ test for H_1^c with $N = 10$ yields error rates $\alpha(x^*) = .377$ and $\beta(x^*) = .306$, the error rates fall to $\alpha(x^*) = .097$ and a $\beta(x^*) = .108$ when $N = 100$, and $\alpha(x^*) = .00003$ and a $\beta(x^*) = .00004$ when $N = 1000$.

FIGURE 13 - Number of Observations and the Nature of the $\alpha + \beta$ Minimizing Test

The relationship between α , β and N can be further illustrated by considering the normal approximations to the null and alternative distributions as N varies. Table 6 lists the mean and standard deviation of the null and alternative distribution, H_1^c of x for $N = 10, 100$, and 1000 . The table also lists the critical values x^C for a one tailed $\alpha = .05$ NHST, and the critical value x^* for the $LR(x^*) = 1$ test that minimizes $\alpha + \beta$.

Table 6 – Distribution Properties and N Using Normal Approximation, with H_1^c as Alternative Hypothesis

N	Null Hypothesis			Alternative Hypothesis		
	Mean	SD	x^C	Mean	SD	x^*
10	5	1.58	2.40	3.75	1.53	4.43
100	50	5	41.78	37.5	4.84	43.71
1000	500	15.81	473.99	375	15.31	436.50

Figure 14A depicts the normal approximation to the null distribution $f(x|H_0)$ and alternative distribution $f(x|H_1^c)$ for $N = 10$. The $\alpha = .05$ one-tail critical value for this distribution is $x^C = 2.40$, and the power of a test that rejects H_0 when $x < x_c$ equals $1 - \beta = .189$. Thus, as demonstrated previously, fixing the significance level at $\alpha = .05$ results in a high Type II error rate ($\beta(x^C) = .811$).

FIGURE 14A – Null and Alternative Distribution H_1^c ($N = 10$)

Increasing the critical value for rejection of the null hypothesis can increase the power of the test and lower the sum of the error rates. The $LR(x^*) = 1$ test that minimizes $\alpha + \beta$ is the test that rejects H_0 when $x < x^* = 4.43$. The figure illustrates the effect on the Type I and Type II errors of moving from the $\alpha = .05$ significance level cutoff of $x^c = 2.40$ to $x^* = 4.43$. The increase in the Type I error rate ($\Delta\alpha = .309$) is more than offset by the decrease in the Type II error rate ($\Delta\beta = -.482$).⁶⁶

However, because of the large overlap between the null and alternative distributions, any test—even the test $LR(x^*) = 1$ that minimizes $\alpha + \beta$ —will have a high total error rate. As a result, even the $LR(x^*) = 1$ test that rejects H_0 for values of $x < x^* = 4.43$ produces Type I and Type II error rates that are both over .3, and the total error rate is just under .7 ($\alpha(x^*) + \beta(x^*) = .688$).⁶⁷

The properties of the normal approximations to the null and alternative sampling distributions when $N = 100$ are depicted in Figure 14B. When $N = 100$, the means of the null and alternative sampling distributions are 2.5 standard deviations apart, reducing the overlap between the two distributions. As a result, the power associated with the fixed α

⁶⁶ Because the null and alternative distributions have different standard deviations, the test that minimizes $\alpha + \beta$ will not equalize the Type I and Type II error rates. See Meyerson & Meyerson, *supra* note 36 at 840 (Advocating a test where $\alpha = \beta$). See also Gelbach, *supra* note 32, who provides conditions sufficient for the result that when the likelihood ratio equals 1, α and β will be equal and less than 0.5 (the conditions include that the alternative hypothesis is specific, as considered in the present paper); these conditions rely on the monotone likelihood ratio property, which need not hold when the distributions under the null and alternative hypotheses have different variances.

⁶⁷ These error rates are based on the normal approximation, and thus differ slightly from the rates based on the binomial probabilities reported above.

= .05 test exceeds .8. The Type I and Type II errors under the $LR(x^*) = 1$ test are smaller, with both approximately equal to .10.⁶⁸

FIGURE 14B - Null and Alternative Distribution H_1^c (N = 100)

Finally, Figure 14C shows the normal approximations to the null and alternative sampling distributions when $N=1000$. The means of the distributions are over seven standard deviations apart, and overlap of the distributions is minute. The fixed- $\alpha = .05$ NHST produces a test with near perfect power. By design, this test still has $\alpha=0.05$, so total error probability is always at least 0.05. By contrast, the $LR(x^*) = 1$ test, which lowers the critical value from x^C to x^* only negligibly reduces power even as it reduces the Type I error rate from .05 nearly to zero. The result is a near-perfect test, with α and β both approximately equal to zero. This example shows that when sample size is large, total error costs can be reduced below what the fixed- α test delivers, precisely because such a test fixes α at a positive level.

⁶⁸ The effect of increasing N can also be seen by examining the standardized test statistic Z discussed in note 60, *supra*. For a binomial sampling distribution for counts, the normal test statistic is $Z = \sqrt{N} \frac{(p-p_0)}{\sqrt{p_0(1-p_0)}}$. This expression shows that increasing N will increase Z for any given difference $p - p_0$. As a result, for large N , even small differences from the null probability p_0 will be statistically significant at “standard” fixed levels of α (e.g., .05)

FIGURE 14C – Null and Alternative Distribution H_1^c (N = 1,000)

D. Relationship of the Mathematical Representation of the Preponderance Standard to the Significance Level

Under the preponderance standard, a court should find for the plaintiff when evidence supporting the plaintiffs' case is more likely than the evidence that supports the defendant's case. Under the assumption of equal error costs, the standard of proof threshold from the preponderance standard is the $LR(x^*) = \pi$ test given in equation (10).

The prior sections have established that the fixed- $\alpha = .05$ NHST is generally not the same as the $LR(x^*) = \pi$ preponderance test. This is true in general for any fixed- α NHST test. When a specific alternative is specified, it is not true that the preponderance standard requires a significance level $\alpha = .50$.⁶⁹ Based on a one-tailed test, such an $\alpha = .50$ test would require that the null hypothesis be rejected when the outcome is below the mean of the null distribution or:

$$\mu_0 = Np_0, \text{ or } \alpha = \int_{-\infty}^{Np_0} f(x|H_0)dx = F(Np_0) = .5 \quad (15)$$

⁶⁹ Cohen, *supra* note 6 at 415-6. See also the text in note 5, *supra* (presenting examples of plaintiff's experts arguing for use of a 50% significance level).

FIGURE 15 – Preponderance Standard and Fixed α Tests

Figure 15 depicts the non-equivalence of the preponderance standard and fixed- α NHST when $N=100$. From condition (15), the threshold for the fixed $\alpha = .50$ null hypothesis is determined by the point where the null distribution function $F(x|H_0)$ intersects the horizontal line at .5. This yields a critical value $x^C = \mu_0 = Np_0 = 50$. In contrast, from (11), the threshold x^* for the preponderance standard with diffuse priors requires that the trier of fact finds for the plaintiff when $LR = \frac{f(x|H_1)}{f(x|H_0)} > LR(x^*) = 1$. This condition holds when $x < x^*$. The preponderance standard threshold is determined by the point where the likelihood ratio LR intersects the horizontal line at 1.0. This criterion yields a critical value $x^* = 43.71 < 50$. Thus, $x^* \neq x^C = \mu_0 = Np_0$.

FIGURE 16 - Preponderance Standard and Fixed α Tests

To provide an alternative view of this point, Figure 16 depicts the critical thresholds from the fixed- $\alpha = .05$ NHST, the $LR(x^*) = 1$ test, and the fixed- $\alpha = .50$ NHST based on normal approximations of the null and H_1^c distributions when $N = 100$. As was demonstrated above when $\pi = \omega = 1$, the preponderance threshold is identical to the $LR(x^*) = 1$ threshold, and will generally not coincide with the fixed $\alpha = .05$ threshold. The latter may lie to the left or right of the former. In Figure 16, the $\alpha = .05$ threshold lies to the left of the $LR^* = 1$ threshold. For similar reasons, the preponderance threshold will not coincide with the $\alpha = .50$ threshold either.⁷⁰ Based on the normal approximation, this outcome can only occur if the null and alternative distributions are identical. When they are not, as in the case depicted in Figure 16, the preponderance standard can never be the same as a fixed $\alpha = .50$ null hypothesis test.⁷¹ Moreover, use of such a test would increase the total error cost rate $\alpha + \beta$ relative to the preponderance standard, raising the Type I error rate by .396 and reducing Type II error rates by .095.

E. Preliminary versus Final Adjudication

Finally, the analysis can be used to model other legal thresholds. Much of the discussion of burdens of proof in the law and economics literature focus on the standard applicable to final adjudication. However, as noted above, the same standards are sometimes applied to adjudicate preliminary issues, such as the burden of production for a *prima facie* case of discrimination examined above.

For example, the analysis can be applied to the pleading standard applicable to scienter under the PSLRA, as interpreted by *Tellabs*.⁷² The Court in *Tellabs* describes this pleading standard in the following way:⁷³

In order to survive a motion to dismiss ... a plaintiff must allege facts from which “a reasonable person would deem the inference of *scienter* cogent and at least as compelling as any opposing inference one could draw from the facts alleged.”

Letting H_1 denote the plaintiff’s hypothesis and H_0 denote the defendant’s hypothesis, a natural interpretation of the *Tellabs* standard is that the likelihood ratio

⁷⁰ Cohen, *supra* note 6 at 415-6, argues that the preponderance standard yields the same critical threshold as the $\alpha = .50$ threshold. Setting $\alpha = .50$ results in a cutoff at the mean of the null distribution, which is consistent with rejecting the null hypothesis when the proportion of females hired $p = x/N$ is less than .5, which Cohen equates with the absolute probability preponderance standard. But this $p = x/N < .5$ is not the preponderance standard set out in equations (10) or (11).

⁷¹ See Gelbach, *supra* note 32.

⁷² *Tellabs*, *supra* note 46.

⁷³ *Id.* at 322 (discussing heightened pleading standard under PSLRA).

$LR = \frac{f(x|H_1)}{f(x|H_0)} > LR(x^{Tellabs}) = 1$. Mathematically, the application of our analytical framework to the *Tellabs* pleading standard would be identical to that used to evaluate the preponderance threshold with diffuse priors.

However, while the formal analysis is the same, the circumstances under which the critical likelihood ratio is applied will likely differ in a preliminary adjudication. Motions to dismiss occur prior to the exchange of information through discovery, and thus are likely to be characterized by sampling distributions with less precision, e.g., greater standard errors, than will be the case during a final adjudication post-discovery.

To illustrate this point, consider again the discrimination example used throughout the paper.⁷⁴ Suppose that at the time of the motion to dismiss, there is a limited number of relevant observations (e.g., $N = 10$) corresponding to the current round of hiring. Discovery allows the consideration of data associated with prior hiring cycles, so that $N > 10$. Under these circumstances, the preliminary adjudication will involve overlapping sampling distributions depicted in Figure 14A where $N = 10$, while final adjudication will involve the conditions depicted in Figures 14B and 14C where the sampling distribution are more precise and thus distinguishable because N is much larger than 10. In terms of the significance level implied by the $LR(x^{PS}) = 1$ critical threshold, application to a sampling distribution with $N = 10$ and relatively large standard errors will result in a much higher implied significance level than the application at summary judgment or final adjudication of the $LR(x^*) = 1$ threshold to sampling distributions with relatively low standard errors (e.g., where $N = 100$ or $N = 1,000$).⁷⁵

V. Conclusion

Models of optimal legal decision rules often use loss functions that seek to minimize the sum of error costs and direct costs. In theory, the choice of a statistical cutoff can reflect a similar calculus, but the standard practice under fixed-significance level NHST does not. Thus, it is not surprising that use of fixed- α NHST will not correspond to existing or optimal legal decision rules. This potential for divergence

⁷⁴ To be sure, the pleading standard for discrimination cases is different from that for cases covered by the PSLRA. As with other cases for which there is no statutory pleading standard, the standard for discrimination cases is set forth by Rule 8(a)(2) as elaborated by the plausibility standard in *Twombly*, *supra* note 44, and filtered through *Iqbal*'s application of Rule 1's edict that the Federal Rules of Civil Procedures govern all civil actions. The Supreme Court has insisted that the plausibility standard is not a probability standard, but many commentators, judges, and lawyers have suggested or acted as if it is difficult to understand how that could be so. This response is surely partly due to the Supreme Court's use of language suggesting that the plaintiff has failed to plead plausibly when an alternative story is at least as compelling as the plaintiff's.

⁷⁵ The application of the $LR^* = 1$ test at the MTD stage will result in high total error costs $\alpha + \beta$, but such a test will likely be associated with relatively low costs C .

between NHST statistical cutoffs and legal decision rules can inform how to reconcile the tools of statistical inference to be consistent with the applicable legal standard. Our analysis shows how reconciling the two can be achieved by replacing fixed significance levels with likelihood ratio tests.