Cornell Law Review

Volume 73
Issue 1 *November 1987*

Article 9

Apples and Oranges: Confidence Coefficients and the Burden of Persuasion

D. H. Kaye

APPLES AND ORANGES: CONFIDENCE COEFFICIENTS AND THE BURDEN OF PERSUASION

D.H. Kaye†

TABLE OF CONTENTS

In his recent article entitled *Confidence in Probability*[1] Professor Neil Cohen seeks to overthrow "the currently accepted probabilistic formulation of the burdens of persuasion."[2] This reigning theory, the probabilistic formulation, is a direct application of the branch of statistics, popular in economics and business, known as Bayesian decision theory ("BDT"). As applied to forensic proof, BDT holds

† Professor of Law and Director, Center for Study of Law, Science, and Technology, Arizona State University School of Law. S.B. 1968, Massachusetts Institute of Technology; A.M. 1969, Harvard University; J.D. 1972, Yale Law School. I am grateful to Dennis Karjala, Dennis Young, Laurence Winer, and especially Mikel Aickin for comments on this paper and for discussions of issues that it addresses.

1 Cohen, *Confidence in Probability: Burdens of Persuasion in a World of Imperfect Knowledge*, 60 N.Y.U. L. REV. 385 (1985).

2 *Id.* at 386. According to one commentator, Professor Cohen "advances a new approach to Bayesianism that not only changes the contours of the debate regarding statistics in proof, but also may point the way to the creation of a unified uncertainty function that may assist in evaluating and integrating the various approaches in the years ahead." Ashford, *Take What You Have Gathered From Coincidence: The Importance of Uncertainty Analysis*, 66 B.U.L. REV. 943, 944-45 (1986).

that, in principle, a verdict for plaintiff is justified if an idealized judge or jury, given the parties' evidence, finds that the probability that plaintiff's story is true exceeds some threshold figure. Thus, the theory has two components: (1) the probability that quantifies the idealized factfinder's partial belief, and (2) the critical number that specifies the minimum degree of belief required under the applicable burden of persuasion. The quantification of the factfinder's partial belief is known as a personal (or subjective) posterior probability. It is "personal" because, upon hearing the same collection of evidence, different persons, having different experiences and background knowledge, will arrive at different values for the probability of the story;[3] it is "posterior" because it is a conditional probability, formed after evidence is produced in court. In contrast, the second component of this model, the threshold to which the posterior probability is compared, is impersonal. It is a number reflecting the relative losses associated with the two possible types of error: a finding for the plaintiff when the defendant's story is true (a false alarm) and a failure to find for the plaintiff when the plaintiff's story is true (a miss). According to BDT, the law has adopted a burden of persuasion that minimizes the expected losses.[4] In civil litigation, where the loss for a false alarm equals the loss for a miss, this criterion leads to the "more-probable-than-not" standard.[5] In this way, BDT seems to provide a pleasing and harmonious interpretation of civil litigation's usual requirement of proof by a preponderance of the evidence.

Cohen finds this analysis unsatisfactory, primarily because of the tenacious problem of naked statistical evidence.[6] Like others before him, Cohen perceives a dissonance between the generally accepted BDT interpretation of the civil preponderance of the evidence standard and a few hypothetical cases in which it is imagined that courts will direct verdicts against a plaintiff whose only evidence is overtly statistical.[7] His examples are familiar and artificial: Dr. L.

---

3 Kaye, *Paradoxes, Gedanken Experiments and the Burden of Proof: A Response to Dr. Cohen's Reply*, 1981 ARIZ. ST. L.J. 635, 643-44.

4 *E.g.*, Kaplan, *Decision Theory and the Factfinding Process*, 20 STAN. L. REV. 1065 (1968).

5 *E.g.*, *id.* at 1072. For extensions and qualifications of this result, see Kaye, *The Limits of the Preponderance of the Evidence Standard: Justifiably Naked Statistical Evidence and Multiple Causation*, 1982 AM. B. FOUND. RES. J. 487.

6 In addition to the many articles on this topic cited in Cohen, *supra* note 1, see Brilmayer, *Second Order Evidence and Bayesian Logic*, 66 B.U.L. REV. 673 (1986) (principles of symbolic logic and arithmetic preclude Bayesian treatment of missing evidence); Lempert, *The New Evidence Scholarship: Analyzing the Process of Proof*, 66 B.U.L. REV. 439 (1986) (defending Bayesian analysis of evidentiary rules); Thomson, *Liability and Individualized Evidence*, 49 LAW & CONTEMP. PROBS., Summer 1986, at 199.

7 For a careful analysis of the meager caselaw on the topic, see Brook, *The Use of Statistical Evidence of Identification in Civil Litigation: Well-Worn Hypotheticals, Real Cases, and*

Jonathan Cohen's Paradox of the Gatecrasher[8] and Professor Laurence Tribe's Blue Bus Case.[9] In the Gatecrasher case, the plaintiff contends that a person entered a rodeo without paying the admission price and relies exclusively on proof that 499 of the spectators paid but 501 did not. In the Blue Bus case, the plaintiff alleges that the Blue Bus Company operated the bus that ran her down, and her sole proof of the company's culpability is the fact that it operates eighty percent of the busses in town. Cohen assumes that the plaintiffs could not prevail in these cases.

In my view, the resolution of the perceived dissonance resides in the often overlooked distinction between justified and unjustified naked statistical evidence,[10] and in the application of a negative, spoliation-like inference or doctrine in the latter situation.[11] For example, in both the Gatecrasher and the Blue Bus cases, the plaintiff's failure to adduce some further evidence appears unjustified, because such evidence should be available to them at little cost. With the aid of some subsidiary arguments,[12] it follows that there are good reasons for not requiring defendants to counter the plaintiffs' limited, statistical showing. In contrast, if the hypotheticals are embellished to make the plaintiffs' reliance on naked statistical evidence appear justified, then it is not at all clear that plaintiffs would or should lose. However, because Cohen adds little to the previous criticism of this theory, not much would be gained by pursuing these points.[13] Furthermore, even if I am correct in maintaining that the naked statistical evidence problem is not a true counterexample to the conventional theory, it could well be that some "new model"[14] of the burden of persuasion would provide "a more accurate, comprehensive concept of forensically determined probabilities."[15]

Controversy, 29 ST. LOUIS U.L.J. 293, 299-305 (1985). *See also* Allen, *Rationality, Mythology and the "Acceptability of Verdicts" Thesis*, 66 B.U.L. REV. 541 (1986). The notion that the law clearly and categorically forbids naked statistical evidence, *see, e.g.*, Brilmayer, *supra* note 6; Nesson, *The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts*, 98 HARV. L. REV. 1357 (1985), is largely a myth. *See* Kaye, *A First Look at "Second-Order Evidence"*, 66 B.U.L. REV. 701 (1986).

8 L.J. COHEN, THE PROVABLE AND THE PROBABLE 74-76 (1977).

9 Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329, 1340-41 (1971).

10 *See* Kaye, *supra* note 5.

11 *See* Lempert, *supra* note 6, at 450-62 (explaining divergences between Bayesian model and trial results by fact that trial factfinding is based on subjective probabilities that may include spoliative inference).

12 *See* Kaye, *supra* note 5; Kaye, *supra* note 7.

13 For a more refined presentation of one of the underlying ideas, see Kaye, *Do We Need a Calculus of Weight to Understand Proof Beyond a Reasonable Doubt?*, 66 B.U.L. REV. 657 (1986) (conditional probability analysis can adequately reflect degree of completeness of evidence that party offers).

14 Cohen, *supra* note 1, at 386, 418, 422.

15 *Id.* at 422.

Accordingly, I shall not dwell on Cohen's reliance on the problem of naked statistical evidence to motivate his theory. Instead, I shall spell out the logic and implications of his analysis, and I shall try to apply a similar analysis to a realistic problem—assessing the evidentiary value of polygraph tests.

In Cohen's view, the posterior probability that a plaintiff's story is correct is a "point estimate" or a "sample statistic" founded on "sample" data.[16] From this perspective, Cohen argues that a determination of whether the burden of persuasion is satisfied turns not merely on the posterior probability, but on "information about the sample size" as well.[17] Stripped of statistical jargon, Cohen seeks to express in a formal way the intuitively plausible idea that the best estimate of the probability of an event may be the same in two different cases, but because of differences in the underlying evidence, the uncertainty associated with the two estimates may be quite different. Thus, if I flip a coin of unknown bias ten times and observe six heads, I might estimate the probability of a head on each independent toss to be 0.6, but I would not be too certain about this value. If I toss the same coin another 999,990 times to find another 599,994 heads, I would estimate with great confidence that the probability of heads on each toss is 0.6. As Cohen explains, statisticians capture this intuitive notion with the technical apparatus of a "confidence interval."[18] Cohen grafts this concept onto BDT by insisting that to satisfy the burden of persuasion, the entire confidence interval, rather than just a single number, must exceed the pertinent threshold. Because the resulting theory relies so heavily on confidence intervals, I shall refer to it as confidence interval theory, or "CIT."

Despite its superficial appeal, CIT is incoherent. In technical terms, it is incoherent because it marries the frequentist's confidence coefficient to a subjectivist's posterior probability. In explaining this assertion, I shall demonstrate that the liaison between a confidence interval and a posterior probability is an unholy union (1) that frequentists and subjectivists alike should shun, and (2) that leads to a conception of the burden of persuasion that yields arbitrary and unjustifiable results.

I undertake this task of criticism to expose and correct a persistent misunderstanding—prevalent in law review articles,[19] trea-

---

16 *Id.* at 398. "Point estimate" is defined as "a single estimated figure, or best guess, based on a sampling of the data." *Id.* at 398 n.76.

17 *Id.*

18 *Id.* at 401.

19 *See, e.g.*, Barnes, *A Common Sense Approach to Understanding Statistical Evidence*, 21 SAN DIEGO L. REV. 809, 832 (1984); Braun, *Quantitative Analysis and the Law: Probability Theory as a Tool of Evidence in Criminal Trials*, 1982 UTAH L. REV. 41, 74; Sprowls, *The Admissibility of Sample Data into a Court of Law: A Case History*, 4 UCLA L. REV. 222 (1957).

tises,[20] casebooks,[21] and judicial opinions[22]—of the relationship between "significance" and "confidence," on the one hand, and the posterior probability and the burden of persuasion, on the other. This misunderstanding has important practical implications. As I have noted elsewhere, in cases involving statistical proof, it can engender a false sense of confidence in the implications of statistical evidence.[23] So too, in this article, I shall discuss how recent reports of the "confidence" that judges or jurors can have in the results of polygraph tests in criminal cases could be misinterpreted. Thus, by explaining my disagreements with Cohen's effort to define "confidence in probability," I hope to promote a deeper understanding within the legal profession of the confidence interval's meaning and relation to the burden of persuasion.

I
THE MEANING OF A CONFIDENCE COEFFICIENT

As Cohen describes his confidence interval theory of the burden of persuasion, it sounds very much like the BDT interpretation: Cohen speaks both of a threshold probability required for a verdict for the plaintiff and of a posterior probability that must exceed this threshold. According to CIT, however, the posterior probability is not a single number as it is in BDT, but is a band whose width reflects the amount of information at the factfinder's disposal. According to Cohen, this interval estimate is a confidence interval of the sort described in virtually all elementary statistics texts, and if this confidence interval for the posterior probability does not lie entirely above the threshold probability, then the burden of persuasion has not been satisfied. This section discusses the link between

---

20 *See, e.g.*, R. WEHMHOEFER, STATISTICS IN LITIGATION 56-57 (1985); D. VINSON & P. ANTHONY, SOCIAL SCIENCE RESEARCH METHODS FOR LITIGATION 129 (1985). The introductory chapters of D. BARNES & J. CONLEY, STATISTICAL EVIDENCE IN LITIGATION, 34 n.4, 81-82 (1986), correctly warn against equating "confidence" with the probability that the null hypothesis is true; however, the book in succeeding chapters repeatedly misconstrues the meaning of "confidence." *Id.* at 108-09, 267, 306.

21 *See, e.g.*, W. LOH, SOCIAL RESEARCH IN THE JUDICIAL PROCESS: CASES, READINGS AND TEXT 410 (1984) (survey research estimates).

22 *See, e.g.*, Vuyanich v. Republic Nat'l Bank, 505 F. Supp. 224 (N.D. Tex. 1980), *reconsidered and adhered to*, 521 F. Supp. 656 (N.D. Tex. 1981), *vacated and remanded*, 723 F.2d 1195 (5th Cir.), *cert. denied*, 469 U.S. 1073 (1984).

23 Kaye, *Statistical Significance and the Burden of Persuasion*, 46 LAW & CONTEMP. PROBS., Autumn 1983, at 13. For illustrations of incorrect statements that might overstate the probability of falsity of the hypothesis favored by the opponent of statistical evidence, see Vasquez v. Hillery, 106 S. Ct. 617, 621 & n.3 (1986) (probability of exclusion of blacks from randomly selected grand juries); Rivera v. City of Wichita Falls, 665 F.2d 531, 545 n.22 (5th Cir. 1982) (inferences of purposeful racial discrimination or disparate racial impact); Craik v. Minnesota State Univ. Bd., 731 F.2d 465, 477 (8th Cir. 1984) (gender discrimination).

a confidence interval and a posterior probability. It shows that this link is tenuous and subtle, and that Cohen's frequentist confidence interval does not determine an interval for a posterior probability.

A. Flipping Coins

CIT, as developed by Cohen, may sound like a minor amendment to BDT, but when Cohen tries to explain how the judge or jury arrives at an interval estimate, he does not use posterior probabilities at all. Consider his paradigmatic example: estimating the probability that a coin will turn up heads on each independent toss.[24] When the data consist of observations showing 26,000 heads in 50,000 tosses, he computes the 95% confidence interval (CI) to be .52 ± .004. When the data show 27 heads in 50 tosses, he computes the 95% CI to be .54 ± .14. Finally, when the data show 51 heads in 100 flips, he computes the 95% CI to be .51 ± .098. Because only the first CI does not cover .5, he concludes that this is the only instance in which "we can state with confidence" that the coin is not fair.[25]

At no point in this example is a posterior probability called into play. The probability of heads on each independent toss of a coin is a parameter of a probability model. Let us call this unknown number π. The value of π determines whether the coin is fair, but does not give the probability that the coin is fair. If $\pi \neq .5$, the coin is not fair, but the extent to which differs from .5 does not determine our degree of belief in the claim that $\pi \neq .5$.

If neither π nor the distance between π and .5 determines subjective confidence in the proposition that the coin is fair, what does? Like many judges and attorneys who first encounter a confidence interval,[26] Cohen seems to think that the 95% coefficient for the CI measures the probability that the parameter is within the CI. He asserts that "a ninety-five percent confidence interval surrounding a point estimate describes a region in which we believe the true value will lie ninety-five percent of the time,"[27] and he claims that for the case of 51 heads out of 100 tosses, the CI of .51 ± .098 "represents the region in which the true value will fall ninety-five percent of the time."[28]

These claims are wrong, but interesting. They exemplify, in an exaggerated way, the tendency of even knowledgeable statisticians to work like frequentists but, when put on the witness stand, to ex-

24 Cohen, *supra* note 1, at 400-03.
25 *Id.* at 401.
26 *See supra* note 23.
27 Cohen, *supra* note 1, at 401.
28 *Id.* at 402.

plain their computations like subjectivists.[29] To clarify my meaning, I must describe with some care the nature of the intervals that Cohen computes. The simplicity of his computations, combined with his omission of a fully specified probability model, disguises the subtlety of the meaning of the computed intervals. Consequently, I shall work through Cohen's coin-toss example, emphasizing the logic behind the procedure.

In Cohen's example, the question to be answered with the help of the statistical evidence is whether the coin is fair. We describe the experiment with a probability model. The outcome of any series of tosses can be characterized by a sample statistic: the proportion of heads. Let p denote this sample proportion; for the previously mentioned outcome of 51 out of 100 heads, p is .51. As before, let π represent the unknown parameter in the probability model: the probability of a head on each toss. Suppose, for instance, that the coin is slightly biased, in that $\pi = .51$. Now imagine that the experiment of tossing the coin 100 times is repeated over and over. For many of these samples of 100 tosses, the proportion p will be .51. For others, the sample proportion will be .52. For still others, it will be .50. Indeed, if we repeat the experiment enough times, the full range of possible values of p, from zero to one, will be observed. But if π is close to .5, values like .50, .51 and .52 will occur much more frequently than extreme values like zero and one. If we know the value of π, probability theory allows us to state the relative frequency of each possible sample proportion. Figure 1, adapted from Cohen's article,[30] shows the frequencies that would be expected over the long run if $\pi = .51$.

However, Figure I is a fantasy. We do not know that π is .51. We only know that in one set of I00 trials, we observed a sample proportion of .51. How can we infer the value of the parameter of the model from this single sample statistic? The usual procedure is to construct a confidence interval. The recipe for the 95% confidence interval for π is simple. Take the sample proportion $p = .51$ as a point estimate for π, and indicate its precision by saying that π could depart from this number by as much as $\pm$ 1.96 standard deviations. The standard deviation is merely a number that measures how much the relative frequencies are scattered about their central

---

29 *Cf.* Aickin, *Issues and Methods in Discrimination Statistics*, in STATISTICAL METHODS IN DISCRIMINATION LITIGATION 159, 161-62 (D. Kaye & M. Aickin eds. 1986) ("Speaking very roughly, most statisticians can be categorized with regard to their attitudes towards probability as being either *frequentists* or *subjectivists*. Some are committed to one of the two philosophies, while others use whichever approach seems to be appropriate to the nature of the problem at hand.").

30 Cohen, *supra* note 1, at 402. To avoid reproducing certain errors, I have labelled the axes differently from Cohen's Figure 1.

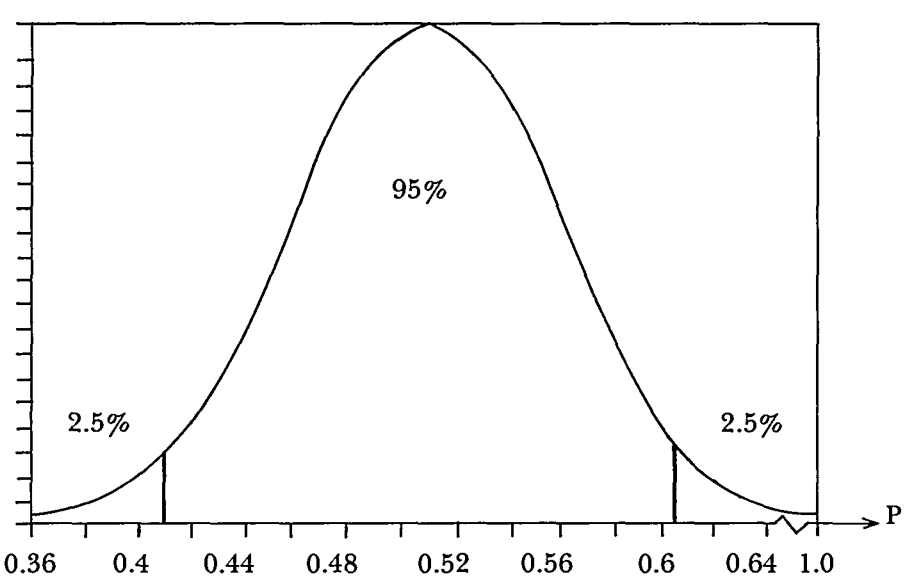

Figure 1.
Theoretical distribution of the random variable p given that the parameter $\pi = .51$.

value π. In this case, the standard deviation is computed according to the formula $\sqrt{p(1-p)/100} = \sqrt{.51(.49)}/10 = .05$.[31] So, the 95% CI is $.51 \pm 1.96(.05) = .51 \pm .098$, as Cohen reports.[32]

Now for the interesting question: where did the 95% and the 1.96 come from? The short answer is that for the kind of numbers we are talking about, the relative frequencies for the possible sample proportions are such that 95% of the sample proportions that would come from repeated experiments lie within $\pm$ 1.96 standard deviations of π. If you think about it enough, you probably can convince yourself that it follows that if we were to use 1.96 in forming the CI, not merely for the estimate based on one sample, but for repeated estimates, then approximately 95% of these CIs would cover the parameter π—whatever π happens to be! Some of these

---

31 Astute readers will have detected a certain sleight of hand here. The standard deviation depends on the central value, π, and we do not know π. We have used our estimate $p = .51$ to compute an estimate for the standard deviation.

32 This common method of calculating the confidence interval gives an approximate solution to the problem of inverting binomial probabilities. The more accurate form of the normal approximation in I. BURR, APPLIED STATISTICAL METHODS 270 (1974), gives a slightly wider interval, $.409 < p < .611$. For still better approximations, see Blyth, *Approximate Binomial Confidence Limits*, 81 J. AM. STATISTICAL A. 843 (1986). From now on, I shall present only the simplest approximations of the exact confidence limits.

CIs would be identical to the one that we computed, .51 ± .098, but in general, they would vary from one sample to the next, both in width and location. Therefore, the confidence coefficient is not the probability that π lies within the lonely interval we observed. Rather, it is the long run frequency with which various and varied CIs would cover the unknown value for π. Confidence pertains not to any specific interval estimate, but to the process for constructing CIs.[33]

Cohen mistakenly concludes that the confidence coefficient gives the probability that π is within the CI because he treats the parameter π, which is an unknown but fixed number, as if it were a random variable. His Figure 1,[34] reproduced below as Figure 2, shows a probability distribution over the "Probability of Heads." Cohen writes that this curve indicates "the probability of various possible true values for the probability of heads for [the] coin."[35] From the frequentist perspective underlying the computation of the CI, however, this picture and these comments make no sense. There is only one possible true value for π, and that value does not vary as we take more samples. What varies is the sample proportion p. The probability distribution in Cohen's figure pertains to p, not to π.[36] He has not drawn a confidence interval at all; unwittingly, he has drawn a prediction interval[37] for p based on an assumed value for π of .51.[38]

B. *High-Tech Supply Company v. Hacker*

Tossing coins is fun, but Cohen also concocts a "more realistic example"[39] to show how CIT applies to forensic proof. High-Tech Supply Company sends a Z99 computer chip to a mail order customer named Hacker. Hacker never receives the chip, but High-Tech sues for the purchase price because the contract called for delivery F.O.B. The dispositive and only disputed issue in the case is whether the chip was functional. The evidence on this issue consists

---

33 *See, e.g.*, G. LUTZ, UNDERSTANDING SOCIAL STATISTICS 315 (1983); L. OTT, AN INTRODUCTION TO STATISTICAL METHODS AND DATA ANALYSIS 74-75 (1977).

34 Cohen, *supra* note 1, at 402.

35 *Id.*

36 As a leading undergraduate statistics text pithily puts it, "The chances are in the sampling procedure, not in the parameter. . . ." D. FREEDMAN, R. PISANI & R. PURVES, STATISTICS 347 (1978).

37 A prediction interval gives the region in which a random variable, such as a sample proportion, is expected to fall a given fraction of the time if that variable is measured repeatedly.

38 For more appropriate diagrams illustrating the meaning of a confidence interval, see, e.g., D. FREEDMAN, R. PISANI & R. PURVES, *supra* note 36, at 349; G. LUTZ, *supra* note 33, at 316.

39 Cohen, *supra* note 1, at 405.

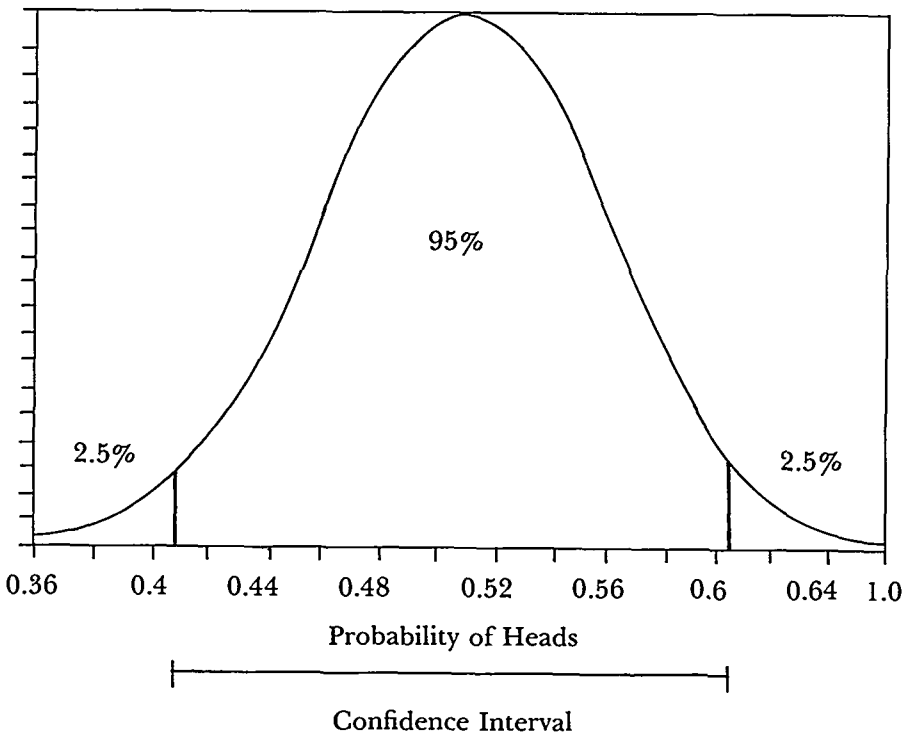

Figure 2.
Cohen's Figure 1, which incorrectly depicts the parameter for the "Probability of Heads" as a random variable.

of High-Tech's testimony that the chip was picked at random from a batch of 1000 chips made by a bankrupt manufacturer and purchased by High-Tech at an auction. High-Tech inspected a random sample of 100 of these chips and found fifty-one to be functional and forty-nine to be defective.[40]

Cohen asserts that the probability that the lost chip was functional is .51, and that a one-sided 95% CI for this probability places it in the interval from .428 to just under 1.0.[41] Because .428 is less than the civil threshold of .5, Cohen concludes that the evidence fails to satisfy the plaintiff's burden of proof.

A more careful statement of the statistical reasoning will clarify

40 Evidently, High-Tech bought the lot after all the chips it tested were replaced.

41 The computation of this CI is slightly troublesome. Cohen assumes that the sample proportion p is normally distributed with variance p(1-p)/100. Because we are sampling a substantial chunk of a finite population without replacement, this assumption is wrong. The distribution of p is hypergeometric with a variance that is smaller by a factor of 900/999. Because Cohen could adjust the numbers in his hypothetical to get the CI that he wants, however, I shall stop quibbling and proceed as if the number of chips in the batch is very large compared to the sample size. With this understanding, the computed CI would apply.

what is really going on here. Let π be the proportion of functional chips in the batch of 1000. π is not a probability, but an unknown parameter. Let f=1 stand for the event that the chip is functional, and f=0 be the event that it is defective. Cohen wants to compute Pr(f=1|p), the probability of randomly drawing a functional chip from a batch of 1000 given the data summarized by the proportion p=.51 of functional chips in the random sample of 100 chips. This probability depends on π. Indeed, it is a particularly simple function of π: the probability of drawing a functional chip at random when there is a proportion π of functional chips is π itself. As in the coin flipping example, the parameter happens to be interpretable as a probability.

So what is π, and hence Pr(f=1|p)? From the standpoint of classical statistics, the one observed value of the sample proportion p = .51 is a point estimate for π. And, if π =.51, then the probability on which the case turns is Pr(f=1|p) =.51. However, a rigorous frequentist cannot say anything about the probability that π is .51, for π is a fixed parameter, not a random variable. Nor does using an interval estimate for π change anything. If the 95% CI is $.428 < \pi < 1$, then approximately 95% of the varied CIs in a long list generated from repeated samples of size 100 would cover the true value of π. What this implies about subjective confidence expressed as a posterior probability that $.428 <$ Pr(f=1|p) < 1 remains obscure.

II
Reforming CIT to Obtain a Consistent Frequentist Theory

We have just seen that Cohen's treatment of a frequentist confidence interval is fundamentally misconceived. Nonetheless, a consistent, purely frequentist variation on CIT is possible in some circumstances. A modification of *High-Tech Supply Company v. Hacker* will illustrate the reasoning. Assume this time that High-Tech did not test a sample before buying a huge batch of chips. To simplify future arithmetic,[42] let us postulate that this batch consisted of one million chips. High-Tech ships one and only one chip, which is intended for Hacker, but is never received by him. High-Tech sues Hacker as before. To prove that it tendered a functional chip, without resorting to prohibitively expensive testing of the remaining 999,999 chips, High-Tech draws a random sample of 100 chips from the pile of 999,999 and counts the number of functional and defec-

42 *See supra* note 41.

tive chips in this sample.[43] Now imagine a long series of such cases, each with its own random sample of 100. If we want to decide the cases by a procedure that would prevent us from concluding that $Pr(f=1|p) > .5$ when $\pi \leq .5$ in 95% or more of the cases, we could do it. We would have to find a value p* such that the probability of observing $p \geq p^*$ given that $\pi \leq .5$ is no larger than .05. This number is $p^* = .58$. If we find for High-Tech only when the sample proportion is at least .58, the rate of false alarms will not exceed 5% if $\pi \leq .5$. To achieve the same result with confidence intervals, we merely form 95% CIs about the sample proportion in each case and find for High-Tech only when these intervals lie entirely above .5. The two procedures are equivalent.[44] For example, when the sample proportion is .51, as posited by Cohen, then the lower bound of the 95% CI for π drops to .428, and we would not find for High-Tech.

So it is possible to use a purely frequentist CIT to keep the conditional probability of false alarms below a specified level, such as .05. This expected conditional error rate is known as a significance level.[45] As in the above example, the significance level is numerically equal to one minus the confidence coefficient.[46]

But what does a confidence coefficient, or its kissing cousin, the significance level, have to do with the burden of persuasion? Cohen recognizes that the connection is "not simple,"[47] and that the 95% coefficient used in every one of his examples may be more demanding than the preponderance-of-the-evidence standard. After rejecting various methods of picking a confidence coefficient, he settles on Dawson's proposal[48] to use a significance level that equalizes the conditional risk of a false alarm and a miss.[49] Using a level that equalizes conditional risks, Cohen believes, will also "equalize the cost of 'wrong' judgments" for plaintiffs and defendants."[50]

This belief, too, is wrong. The cost of each type of error is one

---

43 Please do not ask why High-Tech is introducing this evidence. I am trying to stay within the confines of Cohen's example.

44 For a general statement of the relationship between hypothesis testing and inspecting confidence intervals, see M. DEGROOT, PROBABILITY AND STATISTICS 408-09 (1975); Aickin, *supra* note 29, at 168-69.

45 To avoid confusion, I shall always use the word "level" in conjunction with "significance," and "coefficient" in connection with "confidence."

46 Conceptually, significance and confidence are quite distinct but easily confused. Chandler, *The Statistical Concepts of Confidence and Significance*, 54 PSYCHOLOGICAL BULL. 429 (1957).

47 Cohen, *supra* note 1, at 417.

48 Dawson, *Are Statisticians Being Fair to Employment Discrimination Plaintiffs?*, 21 JURIMETRICS J. 1 (1980); Dawson, *Probabilities and Prejudice in Establishing Statistical Inferences*, 13 JURIMETRICS J. 191 (1973).

49 Cohen, *supra* note 1, at 417.

50 *Id.*

thing, and the conditional probability of these errors is another. The burden of persuasion flows from the former, not the latter. Looking at expected conditional error rates to decide whether data satisfy the preponderance-of-the-evidence standard is like trying to find the shortest path from Oxford to Cambridge by scrutinizing a map of London. To demonstrate that CIT, even with confidence coefficients that equalize expected conditional error rates, is not the road to proof by a preponderance-of-the-evidence, I shall apply CIT to a simplified version of Cohen's *High-Tech* hypothetical. I shall then show that the preponderance-of-the-evidence standard leads to a result that cannot be reconciled with CIT.

A. A Simplified Version of *High-Tech Supply Company v. Hacker*

To facilitate the application of CIT to the hypothetical case of *High-Tech Supply Company v. Hacker*, it is convenient to make a preliminary simplification that restricts the possible values for the proportion of functional chips in the batch of one million. Suppose that the batch came from one of only two machines used in the plant of the bankrupt manufacturer.[51] One machine, the schlock machine, always produces batches of chips of which precisely 50% are functional. The other machine does a little better: it always yields batches in which the proportion of functional chips is precisely 60%. If we knew which batch Hacker's chip came from, the decision would be easy: High-Tech should prevail if the chip sent to Hacker came from a higher quality machine batch ($\pi = .6$), and Hacker should prevail if it came from a schlock batch ($\pi = .5$).

As a first approach to the problem, we may try to reach a conclusion about $\pi = .6$, the hypothesis that the chip came from a better batch.[52] This section considers three methods for reaching such a conclusion: (1) the .05 test mentioned but dismissed by Cohen, (2) the equalized test that he favors, and (3) the traditional more-

---

51 Cohen relies on work that uses similar simplifications to determine the conditional probability of a miss in employment discrimination cases. Cohen, *supra* note 1, at 411. For a suggestion about how this constraint could be generalized, *see supra* note 52. I limit the problem to two deterministic machines to allow the computation of certain conditional probabilities that reveal the disparity between decisions pursuant to CIT and decisions pursuant to the more-probable-than-not standard. The differences between the two approaches remain even in the more complex situations where these probabilities cannot be calculated.

52 This approach is oversimplified in that the posterior probability that $\pi = .6$ is not the posterior probability to which the burden of proof applies. Instead, what must be proven is High-Tech's claim that the randomly selected chip was functional. The probability of this event, computed without knowing the results of the sampling, is some number $Pr(f=1)$. If the chip came from the schlock machine, then $Pr(f=1 \mid \pi=.5) = .5$. If the chip came from the better machine, then $Pr(f=1 \mid \pi=.6) = .6$. Supposing (as we will again in comparing three possible formulations of the burden of persuasion) that the schlock machine produces chips at twice the rate of the better machine, the crucial

probable-than-not test. The analysis demonstrates that the three tests are distinct, and that, contrary to Cohen's assertion, the equalized test does not implement the preponderance-of-the-evidence standard's principle of equal error costs.

1. *The .05 Test*

Confronted with a particular sample proportion p, the factfinder might make one of two decisions. Let D_1 designate the decision that $\pi = .6$, and let D_0 be the decision that $\pi = .5$. Suppose we were to conclude that $\pi = .6$ whenever a sample proportion was at least $p^* = .58$. As indicated above,[53] this is equivalent to a CIT approach with a confidence coefficient of .95. Let us call this decision rule $\delta_{.05}$:

$$\delta_{.05}: D_1 \text{ if } p \geq .58; D_0 \text{ if } p < .58.$$

---

probability would be $Pr(f=1) = Pr(\pi=.5)Pr(f=1 \mid \pi=.5) + Pr(\pi=.6)Pr(f=1 \mid \pi=.6) = (2/3)(.5) + (1/3)(.6) = .53$.

However, the analysis so far makes no use of the sample data p=.51. Because this proportion is more likely to occur when a chip comes from a low quality ($\pi=.5$) batch than a higher quality ($\pi=.6$) batch, it is more likely that the chip came from a schlock batch than our calculations thus far have indicated. Bayes' Theorem reveals that

$$Pr(\pi=.5|p) = \frac{Pr(\pi=.5)Pr(p|\pi=.5)}{Pr(\pi=.5)Pr(p|\pi=.5) + Pr(\pi=.6)Pr(p|\pi=.6)}.$$

The probability of randomly drawing 51 out of 100 functional chips from a large bin in which 50% are functional is $Pr(p=.51 \mid \pi=.5) = .391$, and $Pr(p=.51 \mid \pi=.6) = .183$. Hence, $Pr(\pi=.5 \mid p=.51)$ is .81. Thus, given that the schlock machine grinds out twice as many batches as the better machine, and given that a sample of 100 chips from a batch produced by one of these machines yielded 51 functional chips, we conclude that the probability of the batch having come from the schlock machine is .81.

If the issue in the case were which machine produced the batch (and if no spoliation argument warranted further calculations), then Hacker should prevail. However, as I have indicated, the real issue in the case is whether the chip sent to Hacker was functional. There is, of course, a 50% chance that it was functional even if it came from the schlock machine. Utilizing all of the information, the probability of Hacker having got a functional chip is

$$Pr(f=1|p) = Pr(f=1|\pi=.5)\,Pr(\pi.5|p) + Pr(f=1|\pi=.6)Pr(\pi=.6|p)$$
$$= .5(.81) + .6(.19) = .52.$$

If no spoliation argument is justified, then the fact that $Pr(f=1|p)$ exceeds the civil threshold of .5 implies that the verdict should be for High-Tech.

This analysis obviously is contrived. The two-machine simplification could be replaced with a description of a prior distribution of π in the interval between zero and one. We might visualize this by imagining a vast number of machines of differing quality, or a single machine that wobbles in a stochastically known way from batch to batch.

This simplified analysis nevertheless illustrates some general points. In Cohen's formulation, parameters are never distinguished from probabilities, and it is hard to tell which probability should be compared to the civil threshold of .5. The more complete analysis here makes it clear that the probability that the factfinder must determine is a posterior probability. In our example, it is $Pr(f=1|p=.51)$. This probability depends on prior probabilities related to the manufacturing process and on certain conditional probabilities involving the sample data. In contrast, the confidence interval for π comes from the sample data alone.

53 *Supra* notes 42-44 and accompanying text.

The .05 label is appropriate because if $\pi = .5$, this test errs (in the long run) in 5% of the cases to which it is applied. The .05 significance level is the expected false alarm rate for $\delta_{.05}$ when $\pi = .5$, and it usually is denoted α. Here, $\alpha(\delta_{.05}) = \Pr(D_1 \mid \pi = .5) = .05$.

What is the probability that the test $\delta_{.05}$ will lead to the conclusion that $\pi \leq .5$ when, in reality, $\pi > .5$? This quantity is the expected conditional miss rate for the test. Since the only value in the range $\pi > .5$ that we need to consider is $\pi = .6$, it is just $\Pr(D_0 \mid \pi = .6)$, the probability of a sample proportion $p < .58$ given that $\pi = .6$. We may call this expected miss rate for the $\delta_{.05}$ test (.05). Its value is $\beta(\delta_{.05}) = .34$.

Another way of stating the $\delta_{.05}$ test will prove instructive later. The probability that $p = p^* = .58$ when $\pi = .6$ is .37. The corresponding probability when $\pi = .5$ is .11. The ratio of these probabilities is known as a likelihood ratio. It states how many times more probable the data are to arise when $\pi = .6$ than when $\pi = .5$. The likelihood ratio that corresponds to an observation $p^* = .58$ is

$$LR^* = \frac{\Pr(p=.58 \mid \pi=.6)}{\Pr(p=.58 \mid \pi=.5)} = \frac{.3668}{.1109} = 3.3$$

Thus, to demand that $p > .58$ is to insist on data that are more than 3.3 times more likely to arise if the batch came from the better machine before concluding that the batch in fact came from the better machine. We may restate $\delta_{.05}$ as follows:

$\delta_{.05}$: D_1 if $LR > 3.3$; D_0 otherwise.

2. *The Equalized Test*

Because under the .05 test the risk of a miss is much higher than the risk of a false alarm, Cohen's interpretation of the preponderance-of-the-evidence standard would require an adjustment to p^*. Lowering the value of p^* will make D_1 more probable when $\pi = .5$, thereby increasing the risk of a false alarm in this situation. But it also will make D_0 less probable when $\pi = .6$, decreasing the risk of a miss should $\pi = .6$.[54] At $p^* = .55$, the conditional probability of an error that favors High-Tech is just about equal to the conditional probability of an error that favors Hacker. We may summarize this "equalized test," δ_E, as follows:

δ_E: D_1 if $p > .55$; D_0 otherwise.

---

54 Figure 3 may make this tradeoff easier to visualize. If $\pi = .5$, the sample proportion p has a probability density $f_0(p)$ that is approximately normal with mean .5 and standard deviation .05. If $\pi = .6$, the probability density $f_1(p)$ is approximately normal with mean .6 and standard deviation .0499. The probability α of a false alarm is the area under $f_0(p)$ to the right of p^*. The probability of a miss is the area under $f_1(p)$ to the left of p^*.

The conditional error probabilities for this test are $\alpha(\delta_E.) = .159$ and $\beta(\delta_E.) = .154$.[55] Rephrasing things in terms of confidence intervals, the confidence coefficient is $1 - .159 = .84$. Presumably, in our simplified version of *High-Tech Supply Company v. Hacker*, Cohen would conclude that the preponderance-of-the-evidence standard is not satisfied whenever a one-sided 84% CI about the sample proportion covers $\pi = .5$. Because the observed proportion $p = .51$ is less than p^*, the 84% CI around it would cover $\pi = .5$, and a verdict for Hacker would follow.

Like the .05 test, δ_E may be rephrased in terms of the likelihood ratio exceeding some critical value LR*. Because $LR^* = Pr(p=.55|\pi=.6)/Pr(p=.55|\pi=.5) = .237/.242 = .98$, the equalized test becomes

$$\delta_E: D_1 \text{ if } LR > .98; D_0 \text{ otherwise.}$$

In this particular example, the equalized test works like a maximum-likelihood test.[56] It leads to a finding that $\pi = .6$, and to a verdict for High-Tech, if and only if the data would arise at least almost as frequently when π is .6 as when π is .5.

3. *The More-Probable-Than-Not Test*

Some may find it tempting to reason that if the data are more probable under one hypothesis than another, then the former hypothesis is more likely to be true than the latter. But as a general matter, this reasoning is fallacious. In our hypothetical, the more-

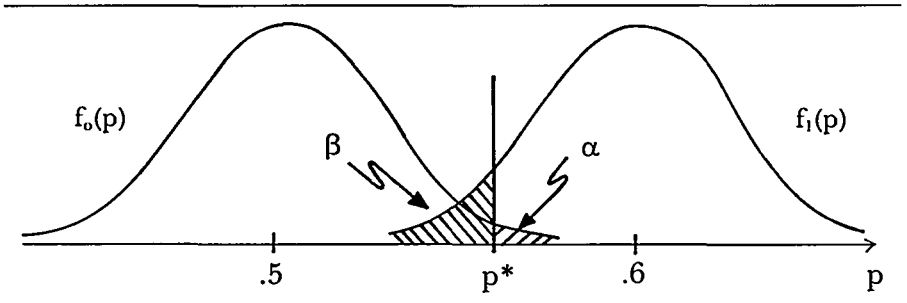

Figure 3. Conditional Error Probabilities for a Decision Rule δ in the Simplified Case of *High-Tech v. Hacker*

55 To equalize these error probabilities more precisely, we would have to choose a p^* slightly larger than .55. Our two digit accuracy here is enough to illustrate the concept.

56 This results from the fact that the distributions of p have the same functional form and almost identical variance under the competing hypotheses $\pi=.5$ and $\pi=.6$. In general, however, an equalized test is not a maximum-likelihood test. Kaye, *Hypothesis Testing in the Courtroom*, in CONTRIBUTIONS TO THE THEORY AND APPLICATIONS OF STATISTICS 331, 344 n.7 (A. Gelfand ed. 1987). Any suggestion to the contrary in Kaye, *supra* note 23, at 23 n.46, is wrong.

probable-than-not standard would usually be phrased in terms of whether the claim that $\pi = .6$ is more probable than the competing claim that $\pi = .5$. If no spoliation argument were justified, we would compute $Pr(\pi = .6|p)$, and, if this figure exceeded .5, we would find that the better machine had been used. Because this decision rule instructs us to accept the claim that has the maximum a posteriori probability, I shall refer to it as the MAP standard.[57] We may express the more-probable-than-not decision rule as

$$\delta_{MAP}: D_1 \text{ if } Pr(\pi=.6|p) > .5; D_0 \text{ otherwise.}$$

To compute $Pr(\pi=.6|p)$, we use Bayes' Theorem which implies that:

$$\frac{Pr(\pi=.6|p)}{Pr(\pi=.5|p)} = LR\frac{Pr(\pi=.6)}{Pr(\pi=.5)}$$

The observed p leads us to update the prior odds in favor of the better machine by multiplying them by a quantity LR, known as the likelihood ratio, to arrive at the posterior odds. LR is computed as the ratio of $Pr(p|\pi=.6)$, the probability of the sample data given a draw from a better batch, to $Pr(p|\pi=.5)$, the probability of the sample data given a draw from the schlock batch.

To facilitate comparison of the δ_{MAP} test with the previous tests, we can express the condition for D_1 in terms of LR. A little algebra reveals that the condition $Pr(\pi=.6|p) > .5$ is fulfilled if and only if the likelihood ratio exceeds $Pr(\pi=.5)/Pr(\pi=.6)$. In other words, δ_{MAP} can be rewritten as follows:

$$\delta_{MAP}: D_1 \text{ if } LR > \frac{Pr(\pi=.5)}{Pr(\pi=.6)}; \quad D_0 \text{ otherwise.}$$

4. *The Tests Compared*

Table 1 summarizes the three decision rules. Compare the expression just given for δ_{MAP} with the LR formulations given earlier for $\delta_{.05}$ and δ_E. In $\delta_{.05}$ and δ_E, the critical LR was determined by looking to the conditional error probabilities α and β for the test. (The .05 test insists that α be kept to .05, leading to the condition LR > 3.33. The equalized test, less demanding, insists on equating α and β, which led to the condition LR > .98.) The δ_{MAP} test, by contrast, does not look to α and β as such. Instead, it requires that the likelihood ratio exceed the prior odds against this event.

Suppose, for example, that the schlock machine churns out

---

57 *See* J. MELSA & D. COHN, DECISION AND ESTIMATION THEORY (1978). The MAP criterion can also be understood as a special case of a Bayes' decision rule. *See supra* text accompanying note 5.

Decision rule (of form D_1 if LR > LR* or p > p*)	LR*	p*
$\delta_{.05}$	3.3	.58
δ_E	.98	.55
δ_{MAP}	$\frac{Pr(\pi=.5)}{Pr(\pi=.6)}$	a function of $\frac{Pr(\pi=.5)}{Pr(\pi=.6)}$

Table 1.
Decision rules for finding $\pi=.6$ given a sample proportion p in the simplified version of *High-Tech Supply Company v. Hacker.*

twice as many batches of chips per day as the better one.[58] Then we might say that the probability that $\pi = .5$ is $Pr(\pi=.5) = 2Pr(\pi=.6)$. Because the chip must have come from one of the two machines, $Pr(\pi=.5) + Pr(\pi=.6) = 1$. Therefore, $Pr(\pi=.5) = 2/3$ and $Pr(\pi=.6) = 1/3$, and the more-probable-than-not standard becomes

$$\delta_{MAP}: D_1 \text{ if } LR > 2; D_0 \text{ otherwise.}$$

The conditional error probabilities for this test are $\alpha(\delta_{MAP}) = .08$ and $\beta(\delta_{MAP}) = .27$.

Comparing the expression for δ_{MAP} to those for $\delta_{.05}$ and δ_E, illustrates that the δ_{MAP} test differs from the other two tests. Depending on the values of $Pr(\pi=.5)/Pr(\pi=.6)$ and p, the equalized test can indicate that $\pi = .6$ even when it is more probable than not that $\pi = .5$.

B. The Burden of Persuasion

I have contended that δ_{MAP} is a formal way of stating the traditional civil burden of persuasion. I believe that this identification of δ_{MAP} with the civil burden of persuasion is accurate because both rules say the same thing about the posterior probability on which cases such as *High-Tech Supply Company v. Hacker* seem to turn. Nevertheless, Cohen seems to feel that δ_E captures a principle of equality implicit in the preponderance of the evidence standard.[59] I have stated that Cohen's view stems from a confusion between the *cost* of

58 *Cf. supra* note 52.

59 *See also* M. FINKELSTEIN, QUANTITATIVE METHODS IN LAW 65-69 (1978). For criticism of Finkelstein's equalization principle, see Kaye, *Naked Statistical Evidence* (Book Review), 89 YALE L.J. 601 (1980).

each type of error and the conditional *probability* of these errors;[60] an explanation of that criticism follows.

The only nonsuperficial analysis of the civil burden of persuasion that I have seen builds on the premise that in civil litigation, a false alarm and a miss are equally serious mistakes.[61] If this premise is correct, we should strive to keep the total probability of these mistakes to a minimum without regard to the direction that they take.[62] To support a rule that increases the expected rate of errors, some other argument besides a vague claim that the civil burden of persuasion has something to do with equating certain error probabilities must be forthcoming.[63] Without such an argument, there is no reason to interpret a "preponderance of the evidence" as evidence at least as powerful as that which equalizes the conditional error probabilities. The decision rule should minimize the total probability of error, without regard to the type of error.

The rule that minimizes total error probability is δ_{MAP}.[64] In our hypothetical, for example, the total probability of error[65] under $\delta_{.05}$ is $Pr(e) = .05Pr(\pi=.5) + .34Pr(\pi=.6) = .05(2/3) + .34(1/3) = .15$. Under δ_E, the probability of error is $.159Pr(\pi=.5) + .154Pr(\pi=.6) = .16$. Under δ_{MAP}, the total probability of error is $.08Pr(\pi=.5) + .27Pr(\pi=.6) = .14$. In this example, the expected error rate for the $\delta_{.05}$ and δ_E rules exceeds the optimal level that the δ_{MAP} rule achieves. Thus, if the underlying principle of the civil burden of persuasion is that false verdicts for plaintiffs and false verdicts for defendants are each to be avoided with equal vigor (so that we should minimize the expected total error rate), then a decision

---

60 *See supra* text following note 50.

61 *E.g.*, Ball, *The Moment of Truth: Probability Theory and Standards of Proof*, 14 VAND. L. REV. 807, 816-18 (1961). Some commentators reject this premise. For instance, Tyree, *Proof and Probability in the Anglo-American Legal System*, 23 JURIMETRICS J. 89, 93 (1982), asserts that where "community standards" treat one type of error as more costly than another, evidence may be excluded to raise, *sub silentio*, the burden of persuasion.

62 *See* Kaye, *supra* note 59. We might consider the expected costs and benefits of obtaining more evidence; in essence, this is what is done by the spoliation analysis mentioned *supra* at text accompanying notes 10-12. Alternatively, these considerations may be incorporated into a more general Bayesian framework that explicitly allows for the possibility of deciding for defendant on the ground that more evidence is needed rather than on the ground that it is more-probable-than-not that, say, $\pi=.5$. Analyses along these lines, as I have said before, are beyond the scope of this article.

63 *Cf.* Kaye, *supra* note 5 (identifying limited circumstances under which it might be reasonable to trade off increased probability of error for better balance of errors).

64 *See* M. DEGROOT, *supra* note 44, at 374-75.

65 The probability of error is a function of the conditional error probabilities α and β and the prior probabilities. Specifically, the risk of a false alarm for a test δ is $Pr(D_1 \mid \pi=.5) = Pr(D_1 \mid \pi=.5)Pr(\pi=.5) = \alpha Pr(\pi=.5)$. Likewise, the risk of a miss is $Pr(D_0 \mid \pi=.6) = Pr(D_0 \mid \pi=.6)Pr(\pi=.6) = \beta Pr(\pi=.6)$. Since false alarms and misses are the only possible errors and are disjoint events, the total error probability is $Pr(e) = Pr(\text{false alarm or miss}) = \alpha Pr(\pi=.5) + \beta Pr(\pi=.6)$.

rule that maximizes a posteriori probability (the MAP standard) is better than one that uses a confidence coefficient or an equalized significance level.[66]

III
"CONFIDENCE" IN POLYGRAPH TESTING

My effort to demonstrate the inability of CIT to capture the essential elements of the burden of persuasion has been more technical than I would have liked. Many legally trained readers may be prone to dismiss as impenetrable a paper peppered with numbers and strewn with symbols. My intent, of course, is to clarify rather than obfuscate. When describing and exploring mathematical models of legal concepts like probative value or the burden of persuasion, ambiguous antecedents are an invitation to disaster, and a consistent and detailed notation is the only practical way to avoid such ambiguities.[67] Applied probability theory is a notoriously slippery field. It is tempting to make intuitively appealing assertions without taking the effort to verify them. It is frightfully easy to speak generally of "probability" or "confidence" when what one means is a specific conditional probability, a parameter in a statistical model, or a likelihood. Thus, the first part of my critique of CIT was that it merely quantifies the sampling error in a statistic. CIT fails to address the possible uncertainty in the posterior probability on which the case should turn, let alone to tell us what do to in the face of such uncertainty.

To illustrate this potential for mischief in a more realistic context, and to identify some sources of uncertainty in arriving at posterior probabilities, it is instructive to consider a recent suggestion by David Raskin, a psychologist who is pre-eminent in polygraph research. Raskin maintains that empirical studies establish a 77% to 92% "confidence" in polygraph testing's capacity to reveal deception by criminal suspects.[68] This "confidence," unlike Cohen's,

---

66 The numbers in this illustration of the expected error minimizing property of δ_{MAP} depend on the prior odds of 2:1 for the schlock machine. If we used a different set of prior odds, δ_{MAP} still would be of the form D_1 if, and only if, $LR > LR^*$, but a different LR^* would apply. This new LR^* would ensure that a finding that the better machine had been the source of the chip would not occur unless the posterior odds favored the better machine. This is what it means to say that δ_{MAP} is a maximum a posteriori test. In contrast, the critical values p^* or LR^* for $\delta_{.05}$ and δ_E do not adjust themselves to the prior odds. Instead, they flow entirely from characteristics of α or β. For some prior odds, $\delta_{.05}$ or δ_E may turn out to reach the same results as δ_{MAP}, but these coincidences do not detract from the claim that only δ_{MAP} is the conceptually appropriate formalization of the preponderance-of-the-evidence standard.

67 *Cf.* Kruskal, *Terms of Reference: Singular Confusion About Multiple Causation*, 15 J. LEGAL STUD. 427 (1986).

68 Raskin, *The Polygraph in 1986: Scientific, Professional and Legal Issues Surrounding Ap-*

does refer to a posterior probability, and this probability can be expressed in terms of confidence intervals. Nonetheless, although these intervals do convey some sense of the imprecision of the estimates of the posterior probability, even these correctly interpreted intervals are not easily related to the burden of persuasion.

To explain what Raskin means by "confidence," I need to introduce some standard notation.[69] A polygraph test, like any medical or psychological test, detects certain symptoms of a disease or condition. Let the class of people with the condition (deception) be D. After administering a polygraph test to a suspect, the examiner reports either that the suspect has the symptoms and belongs to the class (S) or that he or she does not ($\bar{S}$).[70] Two probabilities describe the accuracy of such a test. The probability that a person who is deceptive is classified correctly is known as the sensitivity: $\eta = \Pr(S|D)$. The probability that a person who is not deceptive is classified correctly is known as the specificity: $\theta = \Pr(\bar{S}|\bar{D})$. When the polygraph test is incriminating, the posterior probability of interest is the conditional probability $\Pr(D|S)$ that a person classified as deceptive really is deceptive.[71] In certain contexts, this probability is called the predictive value of a positive test, or "PVP" for short. As we soon shall see, this is what Raskin means by "confidence."

PVP is related to sensitivity and specificity, and the prevalence of deception in the population tested also plays a leading role. Letting the prevalence or base rate be $\beta = \Pr(D)$, and assuming that a person tested can be regarded as randomly selected from that population, Bayes's Theorem can be written in the following form:[72]

$$\text{PVP} = \frac{\beta\eta}{\beta\eta + (1-\beta)(1-\theta)}$$

The PVP figures of .77 and .92 mentioned above come from this formula with two selected sets of estimates for β, η and θ. The estimated values (which we may call b, h, and t) come from the various studies of polygraph accuracy and from certain polygraphers'

---

plications and Acceptance of Polygraph Evidence, 1986 UTAH L. REV. 29, 59-60; Raskin & Kircher, *The Validity of Lykken's Criticisms: Fact or Fancy?*, 27 JURIMETRICS J. 271, 275 (1987).

69 This notation as well as most of the mathematical analysis is taken from Gastwirth, *The Statistical Precision of Medical Screening Procedures: Application to Polygraph and AIDS Antibodies Test Data*, 2 STATISTICAL SCIENCE 213 (1987).

70 For convenience, we ignore the possibility of a report that the test is inconclusive.

71 Likewise, when a test result is exculpatory, the pertinent posterior probability is $\Pr(\bar{D}|\bar{S})$.

72 For a detailed explanation, see Kaye, *Reflections on the Validity of Tests: Caveant Omnes*, 27 JURIMETRICS J. 349 (1987).

experience with field tests of criminal suspects.[73] Because Raskin does not consider the statistical variability in these estimates, his figures for the PVP are point estimates, providing no insight into the sampling error associated with his computations.

Among the many conceivable sources of error in Raskin's figures for PVP, we can estimate the probable extent of the sampling error. One statistician, Joseph Gastwirth, has done so.[74] Sampling error, as we saw earlier, arises from the fact that quantities are estimated on the basis of random samples whose means tend to vary about some central value. In this case, the estimates b, h, and t each come from (presumably) random samples of limited size. Variability in the estimate b of the prevalence β, variability in the estimate h of the sensitivity n, and variability in the estimate t of the specificity θ all affect the precision of the estimate of PVP.[75] For the sample sizes and proportions in question, Gastwirth finds the standard deviation of the estimate of PVP to be around .05.[76] Consequently, the 95% confidence interval[77] for Raskin's lowest estimate of the "confidence" (that is, of PVP), is something like $.77 \pm .10$.

One would imagine that Professor Cohen would approve of this confidence interval treatment of PVP. Indeed, the confidence interval analysis does show that sampling error alone renders somewhat fuzzy the figures that Raskin quotes. We would be deceiving ourselves if we accepted at face value his claim that the lower limit for "confidence" is .77. Broadening the range of our estimate of PVP, however, does little to help us decide whether a positive test result, standing alone, would satisfy the pertinent burden of persuasion; sampling error could raise the estimated PVP as easily as it could lower it. Our best, unbiased estimate of PVP remains the point estimate. If PVP were the posterior probability[78] to be placed alongside some threshold, as Bayesian decision theory prescribes, then it is

---

73 *See* Kaye, *supra* note 72.

74 *See* Gastwirth, *supra* note 69.

75 From the Bayesian perspective (which Gastwirth does not adopt), neither the prior probability (which b estimates) nor the likelihood ratio (which $h/(1-t)$ estimates) is exactly known. Using b for the prior probability raises legal as well as statistical issues. *See* Kaye, *The Polygraph and the PVP*, in 2 STATISTICAL SCIENCE 223 (1987).

76 The .05 figure is derived in Table 2 of Gastwirth, *supra* note 69, which uses slightly different estimates of β, η, and θ than those implicit in the .77 figure of PVP.

77 This confidence interval is formed by moving just under two standard deviations on either side of the estimate for PVP.

78 The PVP is a posterior probability, but it is not a personal probability. The prevalence that plays the role of the prior probability in Bayes' Theorem is the proportion of deceptive people in the population from which the tested suspect is randomly drawn. A more thoroughly subjectivist treatment might take the prior probability as characterizing the individual suspect in light of the unique circumstances of the case. This subjective probability would be modified à la Bayes' Theorem with a likelihood ratio of $\eta/(1-\theta)$ to yield a posterior probability of deception.

still the point estimate of PVP that would seem to be determinative. The uncertainty in PVP affects neither the cost of errors, the expected rate of errors, nor the expected balance of errors resulting from decisions based on PVP.[79]

CONCLUSION

We have come full circle. The appeal of CIT is that it appears to give some structure to the intuition that a probability derived from a broad and firm base of evidence better justifies a decision than a probability derived from a flimsy bit of evidence. Despite the suggestions of Cohen and his followers, however, the frequentist confidence interval is neither "a constructive advance for the Bayesian system" nor a device that "extends to the full range of subjectivist probabilities."[80] On the contrary, there is little breathing space in BDT for the frequentist's idea of "confidence."[81] Even in the context of the one realistic situation we have considered in which a frequentist interval for a posterior probability can be constructed, the relationship between the interval and the burden of persuasion remains mysterious. Yet, there remains an aversion to rushing to judgment on the basis of unnecessarily imprecise estimates of PVP or other posterior probabilities.

I continue to think that this concern can be handled within BDT, perhaps by explicitly including in the set of possible decisions under consideration the option of gathering further information,[82] or perhaps by adjusting the posterior probability to account for the limited base or quality of the underlying evidence.[83] At the same time, nothing that I have said here proves that these approaches are correct or excludes the possibility of some superior analysis. Others may disagree with my preferred approaches, but it is hard to see how CIT can be accepted as a serious model for the burden of persuasion.

---

79 It does bear on the advisability of gathering more information on β, η and θ. A large CI for PVP indicates that doing so might be optimal. BDT is rich enough to include the choice of gathering further information as an option. Analyzing this option within a Bayesian framework takes us back to the literature on naked statistical evidence mentioned in the Introduction, *see supra* notes 6-15 and accompanying text, but no one has yet provided a fully developed treatment along these lines.

80 Ashford, *supra* note 2, at 945.

81 There is a Bayesian version of the confidence interval that advocates of CIT would do well to consider, if only so as to appreciate the dissonance between the frequentist and the Bayesian approaches to statistical inference. For an explanation of Bayesian confidence regions, see, e.g., V. BARNETTE, COMPARATIVE STATISTICAL INFERENCE 198-200 (2d ed. 1982). For an exploration of the connection between frequentist and Bayesian confidence sets, see Meeden & Vardeman, *Bayes and Admissible Set Estimation*, 80 J. AM. STATISTICAL A. 465 (1985).

82 *See supra* note 79.

83 *See* Kaye, *supra* note 13.

Having relentlessly attacked the frequentist conception of "confidence" as a way of thinking about forensic decisionmaking, I should in closing like to allow it some quarter. The frequentist notion can serve as a metaphor to stimulate further work into "Confidence in Probability." Among the more promising technical devices for expressing "confidence" in personal probability, or something like it, are Bayesian sensitivity analysis,[84] second order probabilities,[85] and belief functions.[86] I hope that my effort to spell out Cohen's theory and its implications will challenge Cohen and others who are dissatisfied with BDT to consider these tools and ideas and ultimately to provide a deeper theory of the burden of persuasion.

84 *See* Kass, Comment, 77 J. AM. STATISTICAL A. 347 (1982).

85 *See* Skyrms, *Higher Order Degrees of Belief*, in PROSPECTS FOR PRAGMATISM 109 (D. Mellor ed. 1980).

86 *See* Shafer, *Lindley's Paradox*, 77 J. AM. STATISTICAL A. 325 (1982).