Criminal Law and Statistical Hypothesis Testing*

Michael W. Trosset†

March 26, 1997

Abstract

This essay develops an extended analogy between decisionmaking in criminal law and the Neyman-Pearson formulation of statistical hypothesis testing. Very little background is assumed. The intended audience comprises students in introductory statistics courses and anyone interested in formal models of legal reasoning.

Contents

1 Introduction 2

2 What is Truth? 2

3 Where There’s Smoke There’s Fire 3

4 A Decision-Theoretic Paradigm 5
   4.1 The States of Nature ............................................... 5
   4.2 The Actor ............................................................. 5
   4.3 Innocent Until Proven Guilty ..................................... 6
   4.4 Beyond a Reasonable Doubt ..................................... 7
   4.5 And To a Moral Certainty ...................................... 8

5 Conclusions 9

*Preliminary draft. The author invites comments.
†Department of Statistics—MS 138, Rice University, 6100 Main Street, Houston, TX 77251-1892 (e-mail: trosset@stat.rice.edu).
1 Introduction

This essay is an attempt to synthesize my professional interest in statistical decision theory and my avid—but decidedly amateur—interest in the law. Its purpose is to develop an extended analogy between two modes of rational decisionmaking. One is an idealized model of how juries reach verdicts in criminal trials; the other is the Neyman-Pearson formulation of statistical hypothesis testing, which is widely employed in the scientific community. This analogy is a staple of elementary statistics courses, but we will pursue it in considerably greater depth than have others.

It will become apparent that there are extremely precise parallels between decisionmaking in criminal law and decisionmaking in statistical hypothesis testing. Since neither of these decision-making procedures seem terribly intuitive to the uninitiated, there is considerable pedagogical value in studying both and using each to deepen understanding of the other. To the extent that one can appeal to a popular knowledge of the law when teaching elementary statistics, one renders more accessible a technical subject to which many students have an aversion; to the extent that one can appeal to the rigors of mathematics when studying the law, one continues a long tradition of grounding legal reasoning in the epistemology of natural philosophy.

2 What is Truth?

In The Antichrist, the philosopher Friedrich Nietzsche suggested that the only biblical figure deserving of praise was Pontius Pilate. Like Nietzsche, we take Pilate’s famous question seriously, for the nature of truth depends on the discipline that we study. To illustrate, we contrast the deductive reasoning of theoretical mathematics with the inductive reasoning of empirical science.

Mathematical truth is the correctness of the hypothetical if $A$ then $B$. For example, if one assumes the axioms of Euclid, then one can logically deduce that the sum of the interior angles of a triangle is necessarily $180^\circ$. This truth does not admit uncertainty; however, there are non-Euclidean geometries in which the sum of the interior angles of a triangle does not equal $180^\circ$, so it is only the hypothetical statement if Euclid then $180^\circ$ that is true. If one asks whether or not the interior angles of triangles in the physical world actually do sum to $180^\circ$, then we leave the realm of mathematics and enter the realm of science. One is obliged to physically measure actual triangles and draw inferences from the observed data—to engage in inductive reasoning. The data may be so compelling as to leave the scientist completely convinced of a particular conclusion, but this is not the same kind of certainty that is produced by deductive proof.1

The preceding discussion illustrates an epistemological distinction on which many philosophers have remarked. For example, in An Essay Concerning Human Understanding, John Locke2 distinguished between demonstrative and probabilistic knowledge.3 This essay is concerned with probabilistic knowledge (as well as with using mathematical probability to reason about proba-

---

1. In fact, the widely accepted theory of general relativity posits that our universe is not Euclidean, in which case there is no mathematical guarantee that the interior angles of physical triangles sum to $180^\circ$. Rather, Euclidean geometry is an excellent approximation of the universe on the scale that we usually observe.

2. Locke’s writings profoundly influenced the early history of legal commentary, e.g. Sir Geoffrey Gilbert’s famous 18th century treatise, The Law of Evidence.

3. As demonstration is the showing the agreement or disagreement of two ideas, by the intervention of one or more proofs, which have a constant, immutable, and visible connection one with another; so probability is nothing but the appearance of such an agreement or disagreement, by the intervention of proofs, whose connection is not constant and immutable, or at least is not perceived to be so, but it, or appears for the most part to be so, and is enough to induce the mind to judge the proposition to be true or false, rather than the contrary.” Book IV: Of Knowledge and Probability, Chapter XV: Of Probability, Paragraph 1: Probability is the appearance of agreement upon fallible proofs.
blistic knowledge), and Locke’s further commentary on the various degrees of such knowledge is of particular relevance:

But there being degrees herein, from the very neighborhood of certainty and demonstration, quite down to improbability and unlikeness, even to the confines of impossibility; and also degrees of assent, from full assurance and confidence, quite down to conjecture, doubt, and distrust: I shall come now... in the next place, to consider the several degrees and grounds of probability, and assent or faith.4

3 Where There’s Smoke There’s Fire

Most knowledge is not demonstrative. Reasoning in the face of uncertainty is a common human experience that often assumes the following form: if a phenomenon seems too extraordinary to be a coincidence, then we conclude that the phenomenon was not a coincidence—and our degree of assent to this probabilistic knowledge is commensurate with the degree of improbability of the coincidence. We proceed to consider several examples of this kind of reasoning.

First we consider the familiar example of a criminal trial. For simplicity, we suppose that the defendant has been charged with a single count of pre-meditated murder and that the jury has been instructed to either convict of murder in the first degree or acquit. The defendant had motive, means and opportunity. Furthermore, two types of blood were found at the crime scene. One type was evidently the victim’s. Laboratory tests demonstrated that the other type was not the victim’s, but failed to demonstrate that it was not the defendant’s. What should the jury do?

The evidence used by the prosecution to try to establish a connection between the blood of the defendant and blood found at the crime scene is probabilistic, i.e. circumstantial. It will likely be presented to the jury in the language of mathematics, e.g. “Both blood samples have characteristics x, y and z; yet only 0.5% of the population has such blood.” The defense will argue that this is merely an unfortunate coincidence. The jury must evaluate the evidence and decide whether or not such a coincidence is too extraordinary to be believed, i.e. they must decide if their assent to the proposition that the defendant committed the murder rises to a level of sufficient certainty to convict. Thus, if the combined weight of the evidence against the defendant is a chance of one in ten, then the jury is likely to acquit; if it is a chance of one in a million, then the jury is likely to convict.

Next we consider an example from the realm of science. A recent study5 of termite foraging behavior reached the controversial conclusion that two species of termites compete for scarce food resources. In this study, a site in the Sonoran desert was cleared of dead wood and toilet paper rolls were set out as food sources. The rolls were examined regularly over a period of many weeks and it was observed that only very rarely was a roll infested with both species of termites. Was this just a coincidence or were the two species competing for food?

The scientists constructed a mathematical model of termite foraging behavior under the assumption that the two species forage independently of each other. This model was then used to quantify the probability that infestation patterns such as the one observed arise due to chance. This probability turned out to be just one in many billions—a coincidence far too extraordinary to be dismissed as such—and the researchers concluded that the two species were competing.

Finally, we consider an artificial example that will permit us to scrutinize the precise nature of probabilistic reasoning. Two siblings, a magician (Arlen) and an attorney (Robin) agree to

---

4 Ibid, Paragraph 2: It is to supply the want of knowledge.
resolve their disputed ownership of an Érè painting by tossing a penny. Just as Robin is about
to toss the penny in the air, Arlen suggests that spinning the penny on a table will ensure better
c randomization. Robin assents and spins the penny. As it spins, Arlen calls “Tails!” The penny
comes to rest with Tails facing up and Arlen takes possession of the Érè.

That evening, Robin wonders if she has been had. She decides to perform an experiment.
She spins the same penny on the same table 100 times and observes 70 Tails. It seems to Robin
that perhaps spinning the penny was not entirely fair, but she is reluctant to accuse her brother
of impropriety until she is convinced that the results of her experiment cannot be dismissed as
coincidence. How should she proceed?

Each spin of the penny is an example of a *Bernoulli trial*, i.e., an experiment with two possible
outcomes. It is customary to label the possible outcomes as “success” (to which the number 1 is
assigned) and “failure” (to which the number 0 is assigned). By convention, Heads are successes
and Tails are failures.

Let \( X_i \) denote the number assigned to the outcome of spin \( i \) of Robin’s experiment. Then

\[
Y = \sum_{i=1}^{100} X_i
\]

counts the number of Heads obtained in \( n = 100 \) spins and Robin observed a value of \( y = 30 \).

Supposing the procedure to be fair, Robin had expected to observe (roughly) 50 Heads. Under this
assumption, she would like to calculate the probability of observing a deviation from 50 at least as
large as 50 – 30 = 20, i.e.

\[
P(|Y - 50| \geq 20) = P(Y \leq 30) + P(Y \geq 70).
\]

It seems reasonable to assume that the spins are *independent*, i.e., the outcome of any one spin
is not affected by the outcome(s) of any other(s), and *identically distributed*, i.e., \( p = P(\text{Heads}) \)
does not vary from spin to spin. Under these assumptions, the random variable \( Y \) has a *Binomial
distribution* and

\[
P(Y \leq 30) + P(Y \geq 70) = \sum_{k=0}^{30} \binom{100}{k} p^k (1-p)^{100-k} + \sum_{k=70}^{100} \binom{100}{k} p^k (1-p)^{100-k},
\]

an expression that can be evaluated as accurately as desired for any choice of \( p \). When \( p = .5 \) (the
spin is fair), this probability turns out to be approximately .0001. Different individuals may have
different reactions to this revelation—Robin decides to stab her brother with a scissors.

In each of the preceding examples, a binary decision was based on a level of assent to probabilistic
evidence. At least conceptually, this level can be quantified as a *significance probability*, which we
loosely interpret to mean the probability that chance would produce a coincidence at least as
extraordinary as the phenomenon observed. This begs an obvious question, which we pose now for
subsequent consideration: how small should a significance probability be for one to conclude that
a phenomenon is not a coincidence?

---

*By “fair,” we mean that \( P(\text{Heads}) = P(\text{Tails}) = .5 \).*

*Following standard statistical practice, we denote random variables with uppercase letters and observed values
of random variables with the corresponding lowercase letters.*

*In fact, the probability that spinning a typical penny will produce Heads is roughly 30 percent.*
4 A Decision-Theoretic Paradigm

We proceed to explicate a formal model for the decisions described in Section 3. The formulation that we describe was proposed for statistical hypothesis testing by J. Neyman and E. S. Pearson in the late 1920s and 1930s. Its applicability to criminal law has been widely noted in introductory statistics courses, but has not (so far as we are aware) been explored in depth.

4.1 The States of Nature

The states of nature are the possible mechanisms that might have produced the observed phenomenon. In the penny-spinning example of Section 3, the states of nature are the Binomial distributions indexed by \( n = 100 \) and \( p \in [0, 1] \). For hypothesis testing, the states of nature are partitioned into two sets or hypotheses. In the penny-spinning example, the hypotheses are \( p = .5 \) (the procedure is fair) and \( p \neq .5 \) (the procedure is not fair); in the legal example, the hypotheses are that the defendant did commit the murder and that the defendant did not commit the murder.

The purpose of hypothesis testing is to decide which hypothesis is correct, i.e., which hypothesis contains the true state of nature. In the penny-spinning example, Robin wants to determine whether or not penny-spinning is fair; in the termite example, Jones and Troset wanted to determine whether or not the termites were foraging independently. The legal example is more interesting to contemplate. Why do we have criminal trials? Indeed, why do we have a criminal justice system? The answer is provided by the United States Supreme Court in *Burlington v. Missouri* (1981):

> Underlying the question of guilt or innocence is an objective truth: the defendant did or did not commit the crime. From the time an accused is first suspected to the time the decision on guilt or innocence is made, our system is designed to enable the trier of fact to discover that truth.

Thus, in our legal example, we consider the purpose of the trial to be the rendering of a decision as to whether the defendant did or did not commit the murder.

It is interesting to note that Vincent Bugliosi cites the preceding passage in support of his contention that the legal community’s grasp of the distinction between "proving guilt beyond a reasonable doubt" and "determining guilt or innocence" is "fuzzy, unarticulated, visceral, and not sufficiently conceptualized in their minds to enable or compel them to speak or write correctly on the subject." We endeavor to do better, arguing that the Neyman-Pearson formulation of statistical hypothesis testing provides an ideal paradigm in which to clarify these issues. Accordingly, we emphasize that the hypotheses about the states of nature are hypotheses about factual guilt or innocence. In subsequent subsections we will see that "proving guilt beyond a reasonable doubt" can be precisely articulated by describing how the jury reaches (or is supposed to reach) its decision.

4.2 The Actor

The states of nature having been partitioned into two hypotheses, it is necessary for a decisionmaker (the actor) to choose between them. In the penny-spinning example, the actor is Robin; in the

---

9It can be argued quite persuasively that hypothesis testing is not appropriate in either situation. For example, should Robin feel cheated if it turns out that the true probability of Heads is \( p = .49999 \) rather than \( p = .50000 \)? Many statisticians feel that the scientific community overuses hypothesis testing and that other forms of statistical inference are often more appropriate. Indeed, it is a rather preposterous conceit to imagine that any single scientific article should endeavor to decide whether a theory is or is not true. A more sensible enterprise is to simply set forth the evidence discovered by the authors, e.g., by constructing confidence sets, and allow evidence to accumulate until the scientific community reaches a consensus.

Termite example, the actor is the team of researchers; in the legal example, the actor is the jury.

Statisticians often describe hypothesis testing as a game that they play against nature. To study this game in greater detail, it becomes necessary to distinguish between the two hypotheses under consideration. In each example, we declare one hypothesis to be the null hypothesis ($H_0$) and the other to be the alternative hypothesis ($H_1$). The logic for determining which hypothesis is $H_0$ and which is $H_1$ is not yet apparent, but suppose that we have declared the following null hypotheses: (1) $H_0$: the defendant did not commit the murder, (2) $H_0$: the termites are foraging randomly, and (3) $H_0$: spinning the penny is fair. Having done so, the game takes the following form:

<table>
<thead>
<tr>
<th>State of Nature</th>
<th>$H_0$</th>
<th>$H_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actor's Choice</td>
<td>$H_0$</td>
<td>Type I error</td>
</tr>
<tr>
<td></td>
<td>$H_1$</td>
<td>Type II error</td>
</tr>
</tbody>
</table>

There are four possible outcomes to the game, two of which are favorable and two of which are unfavorable. If the actor chooses $H_1$ when in fact $H_0$ is true, then we say that a Type I error has been committed. If the actor chooses $H_0$ when in fact $H_1$ is true, then we say that a Type II error has been committed. In the context of a criminal trial, a Type I error occurs when a jury convicts a factually innocent defendant and a Type II error occurs when a jury acquits a factually guilty defendant.

### 4.3 Innocent Until Proven Guilty

Because we are concerned with probabilistic evidence, any decision procedure that we devise will occasionally result in error. Obviously, we would like to devise procedures that minimize the probabilities of committing errors. Unfortunately, there is an inevitable tradeoff between Type I and Type II error that precludes simultaneously minimizing the probabilities of both types.\(^{11}\) The distinguishing feature of hypothesis testing (and Anglo-American criminal law) is the manner in which it addresses this tradeoff.

The Neyman-Pearson formulation of hypothesis testing accords the null hypothesis a privileged status: $H_0$ will be maintained unless there is compelling evidence against it.\(^{12}\) In the penny-spinning example, Robin considered a significance probability of $p = .0001$ to be compelling evidence against the null hypothesis that penny-spinning is fair. In the termite example, Jones and Trosset considered an extremely small significance probability to be compelling evidence against the null hypothesis that two termite species forage independently. In a criminal trial, the principle of according the null hypothesis a privileged status has a familiar characterization: the defendant is “innocent until proven guilty.”

According the null hypothesis a privileged status is equivalent to declaring Type I errors to

---

\(^{11}\)To see this, consider two juries. The first jury always acquits and the second jury always convicts. Then the first jury never commits a Type I error and the second jury never commits a Type II error. The only way to simultaneously better both juries is to never commit an error of either type, which is impossible with probabilistic evidence.

\(^{12}\)It is instructive to contrast the asymmetry of the Neyman-Pearson formulation with situations in which neither hypothesis is privileged. In statistics, this is the problem of determining which hypothesis better explains the data. This is discrimination, not hypothesis testing. In law, this is the problem of determining whether the defendant or the plaintiff has the stronger case. This is the criterion in civil suits, not in criminal trials.
be more egregious than Type II errors. This preference can sometimes be glimpsed in scientific applications (because science is conservative, it is generally considered more egregious to falsely reject than to falsely accept the prevailing wisdom that termite species forage independently), but it is an article of faith in criminal law. In his famous Commentaries, William Blackstone opined that “it is better that ten guilty persons escape, than that one innocent man suffer;” and in his influential Practical Treatise on the Law of Evidence (1824), Thomas Starkie suggested that “The maxim of the law…is that it is better that ninety-nine…offenders shall escape than that one innocent man be condemned.” In Reasonable Doubts (1996), Alan Dershowitz quotes both maxims and notes anecdotal evidence that jurors actually do prefer committing Type II to Type I errors: on Prime Time Live (October 4, 1995), Simpson juror Anise Aschenbach stated, “If we made a mistake, I would rather it be a mistake on the side of a person’s innocence than the other way.”

4.4 Beyond a Reasonable Doubt

To actualize its antipathy to Type I errors, the Neyman-Pearson formulation imposes an upper bound on the maximal probability of Type I error that will be tolerated. This bound is the significance level, conventionally denoted \( \alpha \). The significance level is specified (prior to examining the data) and only decision rules for which the probability of Type I error is no greater than \( \alpha \) are considered.\(^{14}\)

To fix ideas, we consider the penny-tossing example and specify a significance level of \( \alpha \). Let \( P \) denote the significance probability that results from performing the analysis in Section 3 and consider a rule that rejects the null hypothesis \( H_0 : p = .5 \) if and only if \( P \leq \alpha \). Then a Type I error occurs if and only if \( p = .5 \) and we observe \( y \) such that \( P(|Y - 50| \geq |y - 50|) \leq \alpha \). But the probability of observing such a \( y \) is just \( \alpha \), so we have constructed a level \( \alpha \) test. Thus, \( \alpha \) quantifies the level of assent that we require to risk rejecting \( H_0 \), i.e. the significance level tells us how small a significance probability is required in order to conclude that a phenomenon is not a coincidence.

In statistics, the significance level \( \alpha \) is a number in the continuum \([0, 1]\). It is not possible to quantitatively specify the level of assent required for a jury to risk convicting an innocent defendant, but the legal principle is identical: in a criminal trial, the operative significance level is beyond a reasonable doubt. Starkie (1824) described the possible interpretations of this phrase in Lockean terms:

Evidence which satisfied the minds of the jury of the truth of the fact in dispute, to the entire exclusion of every reasonable doubt, constitute full proof of the fact. …Even the most direct evidence can produce nothing more than such a high degree of probability as amounts to moral certainty. From the highest it may decline, by an infinite number of gradations, until it produces in the mind nothing more than a preponderance of assent in favour of the particular fact.

The gradations are not intrinsically numeric, but it is evident that the problem of defining reasonable doubt in criminal law is the problem of specifying a significance level in statistical hypothesis testing.

In both criminal law and statistical hypothesis testing, the actions typically are described in language that acknowledges the privileged status of the null hypothesis and emphasizes that the

\(^{13}\)This connection was eloquently articulated by Justice John Harlan in a 1970 Supreme Court decision: “If, for example, the standard of proof for a criminal trial were a preponderance of the evidence rather than proof beyond a reasonable doubt, there would be a smaller risk of factual errors that result in freeing guilty persons, but a far greater risk of factual errors that result in convicting the innocent.” See Section 4.4 for a discussion of the phrase “beyond a reasonable doubt.”

\(^{14}\)Such tests are called level \( \alpha \) tests. Ideally, one tries to determine a decision rule in the class of level \( \alpha \) tests that minimizes the probability of committing a Type II error.
decision criterion is based on the probability of committing a Type I error. In describing the action of choosing \( H_0 \), many statisticians prefer the phrase “fail to reject the null hypothesis” to the less awkward “accept the null hypothesis” because choosing \( H_0 \) does not imply an affirmation that \( H_0 \) is correct, only that the level of evidence against \( H_0 \) is not sufficiently compelling to warrant its rejection at significance level \( \alpha \). In precise analogy, juries render verdicts of “not guilty” rather than “innocent” because acquittal does not imply an affirmation that the defendant did not commit the crime, only that the level of evidence against the defendant’s innocence was not beyond a reasonable doubt.\(^{15}\)

4.5 And To a Moral Certainty

The Neyman-Pearson formulation of statistical hypothesis testing is a mathematical abstraction. Part of its generality derives from its ability to accommodate any specified significance level. As a practical matter, however, \( \alpha \) must be specified and we now ask how to do so.

In science, the answer is somewhat irritating. In his extremely influential *Statistical Methods for Research Workers* (1925), Ronald Fisher\(^{16}\) suggested that \( \alpha = 0.05 \) and \( \alpha = 0.01 \) are often appropriate significance levels. These were intended as practical guidelines, but they have become enshrined (especially \( \alpha = 0.05 \)) in the minds of many scientists as a sort of Delphic determination of whether or not a hypothesized theory is true. While some degree of conformity is desirable (it inhibits a researcher from choosing—after the fact—a significance level that will permit rejecting the null hypothesis in favor of the alternative in which s/he may be invested), many statisticians are disturbed by the scientific community’s slavish devotion to a single standard and by its often uncritical interpretation of the resulting conclusions.\(^17\)

The imposition of an arbitrary standard like \( \alpha = 0.05 \) is possible because of the precision with which mathematics allows hypothesis testing to be formulated. Applying this precision to legal paradigms reveals the issues with great clarity, but is of little practical value when specifying a significance level, i.e., when trying to define the meaning of “beyond a reasonable doubt.” Indeed, Alan Dershowitz notes with some frustration a modern tendency to avoid the issue entirely:

> The U.S. Supreme Court, in an act of abject intellectual cowardice, has declared that the term ‘reasonable doubt’ is self-explanatory and, essentially, incapable of further definition. “Attempts to explain the term ‘reasonable doubt’ do not usually result in making it any clearer to the minds of the jury,” the Court has declared, which brings to mind Talleyrand’s quip that “if we go on explaining, we shall cease to understand one another.” Judge Jon Newman of the U.S. Court of Appeals for the Second Circuit recently criticized this approach as follows: “I find it rather unsettling that we are using a formulation that we believe will become less clear the more we explain it.”\(^18\)

Nevertheless, legal scholars have endeavored for centuries to position “beyond a reasonable doubt” along the infinite gradations of assent that correspond to the continuum \([0,1]\) from which \( \alpha \) is selected. The phrase “beyond a reasonable doubt” is still often connected to the archaic phrase

---

\(^{15}\)In contrast, Scottish law permits a jury to return a verdict of “not proven,” thereby reserving a verdict of “not guilty” to affirm a defendant’s innocence.

\(^{16}\)Sir Ronald Fisher is widely regarded as the single most important figure in the history of statistics. It should be noted that he did not subscribe to all of the particulars of the Neyman-Pearson formulation of hypothesis testing. His fundamental objection to it, that it may not be possible to fully specify the alternative hypothesis, does not impact our development, since the alternative hypothesis is specified in a criminal trial.

\(^{17}\)See, for example, J. Cohen (1994). The world is round \((p < .05)\). *American Psychologist*, 49:997–1003.

“to a moral certainty.” This connection survived because moral certainty was actually a significance level, intended to invoke an enormous body of scholarly writings and specify a level of assent:

Throughout this development two ideas to be conveyed to the jury have been central. The first idea is that there are two realms of human knowledge. In one it is possible to obtain the absolute certainty of mathematical demonstration, as when we say that the square of the hypotenuse is equal to the sum of the squares of the other two sides of a right triangle. In the other, which is the empirical realm of events, absolute certainty of this kind is not possible. The second idea is that, in this realm of events, just because absolute certainty is not possible, we ought not to treat everything as merely a guess or a matter of opinion. Instead, in this realm there are levels of certainty, and we reach higher levels of certainty as the quantity and quality of the evidence available to us increase. The highest level of certainty in this empirical realm in which no absolute certainty is possible is what traditionally was called “moral certainty,” a certainty which there was no reason to doubt.  

Although it is rarely (if ever) possible to quantify a juror’s level of assent, those comfortable with statistical hypothesis testing may be inclined to wonder what values of α correspond to conventional interpretations of reasonable doubt. If a juror believes that there is a 5 percent probability that chance alone could have produced the circumstantial evidence presented against a defendant accused of premeditated murder, is the juror’s level of assent beyond a reasonable doubt and to a moral certainty? One would hope not. We may be willing to tolerate a 5 percent probability of a Type I error when studying termite foraging behavior, but the analogous prospect of a 5 percent probability of wrongly convicting a factually innocent defendant is abhorrent.

Unfortunately, little is known about how anyone in the legal system quantifies reasonable doubt. Mary Gray cites a 1962 Swedish case in which a judge trying an overtime parking case explicitly ruled that a significance probability of 1/20, 736 was beyond reasonable doubt but that a significance probability of 1/144 was not. In contrast, Alan Dershowitz relates a provocative classroom exercise in which his students preferred to acquit in one scenario with a significance probability of 10 percent and to convict in an analogous scenario with a significance probability of 15 percent.

5 Conclusions

One of the dominant themes of Barbara Shapiro’s work on the history of the reasonable doubt standard is that “most, if not all, legal scholars were devoted to showing that the standards of evidence and proof in the law conformed to those in other forms of inquiry.” Statistical decision theory, specifically the Neyman-Pearson formulation of statistical hypothesis testing, provides a natural paradigm in which to organize this discourse.

---


20 This discrepancy illustrates that the consequences of committing a Type I error influence the choice of a significance level. The consequences of Jones and Trosset wrongly concluding that termite species compete are not commensurate with the consequences of wrongly imprisoning a factually innocent citizen.

21 M.W. Gray (1983). Statistics and the law. Mathematics Magazine, 56: 67–81. In writing the present essay at Rice University, I cannot resist quoting another of Gray’s examples of statistics-as-evidence: “In another case, that of millionaire W. M. Rice, the signature on his will was disputed, and the will was declared a forgery on the basis of probability evidence. As a result, the fortune of Rice went to found Rice Institute.”


Most students enrolled in elementary statistics courses begin with a greater appreciation of the mechanics of the criminal justice system than of the mathematics of statistical hypothesis testing. One can capitalize on this appreciation to make their introduction to hypothesis testing more intuitive than might be otherwise. However, the analogy between criminal law and hypothesis testing is sufficiently precise that mastering the latter can also provide more rigorous ways of thinking about the former.

Acknowledgments

This essay is dedicated to the memory of Phyllis Barton, my high school debate coach. It evolved from classroom lectures at the University of Arizona and an invited presentation to Richard Tapia’s *Spend a Summer with a Scientist* program at Rice University. Along the way, I greatly benefitted from conversations with James Fleissner, Susan Hall, Barbara Sands, Holly Teeters and James Thompson. I finally was motivated to write an essay in order to respond to arguments in the popular writings of Vincent Bugliosi and Alan Dershowitz that the search for truth in criminal law is fundamentally different from the search for truth in other disciplines. At the time of this preliminary draft, I have been unable to explore all of the primary sources that I have discovered. Thus, I owe a special debt to secondary sources, especially Dershowitz’s *Reasonable Doubts* and Barbara Shapiro’s “Beyond Reasonable Doubt” and “Probable Cause”.