Convergent Evolution in Law and Science: The Structure of Decision-making Under Uncertainty

Michael J Saks, Arizona State University

Available at: https://works.bepress.com/michael_saks/3/
Convergent Evolution in Law and Science: The Structure of Decision-making Under Uncertainty

Michael J. Saks & Samantha L. Neufeld

ABSTRACT

The formal structure of decision-making under uncertainty used in legal trials bears a remarkable similarity to the structure of decision-making under uncertainty used in hypothesis testing in empirical science. The first purpose of this article is to explicate those similarities. Secondly, the article reviews the historical origins of these decision-making schemes in both law and science, finding that they evolved independently of each other to serve similar functions.

INTRODUCTION

Discussions of the nature of law and science typically emphasize their differences. The present article is an exception. We explicate a remarkable similarity of thought that has evolved in these two different worlds when they confronted similar needs to deal with uncertainty. Indeed, to characterize these as "similar" is to understate the parallels. In an area critical to each – decision-making under uncertainty – law and science have recognized, formalized, and adopted approaches that are identical in every important respect, though they use completely different language to describe the same things.

Each field is generally quite familiar with its own concepts of decision-making under uncertainty, but we suspect that few members of one field appreciate the parallels to be found in the other field. We begin with what is most familiar in law, and then turn to its scientific twin. In addition, we reflect on some of the history that brought the two fields to their highly similar states with regard to trials and hypothesis testing.

---

1 See, e.g., David L. Faigman, Legal Alchemy (1999); Steven Goldberg, Culture Clash: Law and Science in America (1996); Susan Haack, Irreconcilable Differences? The Troubled Marriage of Science and Law, 72 Law and Contemporary Problems 1 (2009). Discussions of law and science often address such issues as the role of positive versus normative inquiry, central versus secondary role of truth-seeking, importance of established belief versus search for ideas that are better than what went before, seemingly irreconcilable tensions when they interact (regarding law's reliance on scientific knowledge as evidence as well as the legal regulation of scientific inquiry or application), and other issues.
I. DECISION-MAKING UNDER UNCERTAINTY

A. The Structure of Decision-Making Under Uncertainty in Legal Trials

The decision choices in trials, both criminal and civil, can be represented by the 2x2 matrix in Figure 1. For convenience, let's focus on a criminal trial. At the conclusion of the trial, the factfinder must choose a verdict of guilty or not guilty on each count charged (or on lesser included charges). In reality, the defendant is guilty or is innocent, but that reality can never truly be known to the factfinder. The factfinder can know only the evidence that has been presented during the trial and the inferences the factfinder has drawn from that evidence. The law appreciates that although a defendant can, in some ultimate reality, be innocent, the factfinder can never declare a defendant to be innocent, but is in a position only to declare that the evidence adduced did cross the requisite threshold to guilt (and the defendant is therefore "guilty") or that the evidence did not cross the threshold (and the defendant is therefore "not guilty").

There are two ways a factfinder can be correct: the defendant can be guilty and the factfinder decides to convict; or the defendant can be innocent and the factfinder decides to acquit. And there are two ways the factfinder can be incorrect: the defendant can be guilty but the factfinder decides to acquit; or the defendant can be innocent but the factfinder decides to convict.

The impossibility of declaring "innocence" stems from the fact that in making decisions under uncertainty, one must adopt a no-information starting point and evaluate the extent to which the evidence moves the decision away from that starting point and along the scale toward guilt. The law's familiar starting point in criminal trials is the presumption of innocence. Innocence in this sense is an assumption, a temporary stance, not a finding, never a conclusion. If insufficient evidence of guilt is presented, the factfinders can declare the defendant to be "not guilty," but they cannot know and therefore cannot declare the defendant to be "innocent." Put succinctly, the law recognizes that non-guilt is not the same as innocence. The verdict does not move from not guilty unless and until the evidence of guilt accumulates to a sufficient extent that the no-guilt starting point can be rejected in favor of a finding of "guilty."

Another decision tool is needed: a threshold which, if crossed by the evidence, changes the verdict from the non-guilty presumption to a verdict of guilty. For the law, that is the standard of proof. In criminal cases, that standard is the threshold of reasonable doubt. Factfinders are to deliver a verdict of "guilty" if and only if the evidence and inferences from that evidence are sufficient to cross the threshold of reasonable doubt. In civil cases, of course, the threshold is set lower (preponderance of the evidence, or "the greater weight of the evidence"),

---

2 How does the notion of "actual innocence" fit into this scheme? See, e.g., BARRY SCHECK AND PETER NEUFELD, ACTUAL INNOCENCE (2000); Brandon Garrett, Judging Innocence, 108 COLUM. L. REV. 55 (2008). Actual innocence is not part of the normal decision machinery of the law. Innocence exists only as a presumption which is to be replaced by a decision of either guilt or non-guilt, but not into a finding of "innocence." During post-conviction proceedings, once sufficient evidence accumulates refuting an earlier verdict of a defendant's guilt – such as DNA evidence excluding a defendant from the crime scene and implicating a different individual – courts do set aside convictions on the basis of something akin to actual innocence. But that is a departure from normal trial decision-making.
where the presumption of non-liability favoring the defendant may be rejected on a mere tipping of the balance in favor the plaintiff. In certain other kinds of cases an intermediate threshold (clear and convincing evidence) must be crossed in order to reject the presumption of non-liability.

The law is not indifferent to the two kinds of errors described above. It recognizes that the two kinds of error stand in some ratio to one another. And by changing the standard of proof, the law is changing the ratio of the two kinds of errors.3 Erroneous convictions have a different value to the law than erroneous acquittals. One might say they have disutilities of different magnitudes. Or, most simply, that one type of error is more serious, threatens our values more than the other type of error. Most readers will be familiar with Blackstone’s maxim that, “It is better that ten guilty persons escape than that one innocent suffer.”4 Taken literally, this suggests that the law prefers ten or more erroneous acquittals to one erroneous conviction, though perhaps Blackstone’s comment is more figurative, and represents only a concept, a value. Either way, the point remains: The law regards erroneous convictions as having far more disutility than erroneous acquittals.5 Other legal cultures, quite different from the Anglo-American system, have developed similar notions of the need to balance erroneous convictions and erroneous acquittals.6

This system of making difficult and consequential decisions in the face of uncertainty is no less sophisticated than the one used by scientists and statisticians. In fact, at bottom, the two are twins.

B. The Structure of Decision-Making Under Uncertainty in Science

The decision choices in conducting scientific research can be represented by the 2x2 matrix in Figure 2. At the end of an experiment, the researcher must decide whether to reject the null hypothesis or not reject the null hypothesis. For example, the research might be testing whether a drug causes a harmful side effect or not. In reality, the drug does or does not cause harm, but that reality cannot truly be known to the researcher. The researcher can only know the evidence that has been developed through the study and the inferences that can be drawn from

---


4 WILLIAM BLACKSTONE, COMMENTARIES, Book 4, chapter 27 (1765-1769).

5 This also reveals that the law is not unaware of the risk – indeed, the inevitability – of error, in its procedures. See D. Michael Risinger, Innocents Convicted: An Empirically Justified Factual Wrongful Conviction Rate, 97 J. CRIM. L. & CRIMINOLOGY 761 (2006-2007).

6 Other ratios have been suggested. Hale in 1736 suggested 5:1 (“... it is better five guilty persons should escape unpunished, than one innocent person should die”). Fortescue, writing in 1616, suggested 20:1 (“Indeede I would rather wish twentie euill [evil] doers to escape death through pittie, then one man to bee vniustly [unjustly] condemned”). Benjamin Franklin proposed 100:1. Maimonides, in the 12th Century, argued for 1000:1. In Genesis, God agrees to spare the city of Sodom if 50 righteous can be found there, and Abraham persuades God to change that to 10 (either of them being a ratio of (perhaps) hundreds or even thousands to 1). In a rather comprehensive review of court rulings, Biblical references, historical views, and other sources, Alexander Volokh, n Guilty Men, 146 UNIV. PENNSYLVANIA L. REV. 173 (1997), found the ratio to range from 1:1 to 5000:1.
that evidence. Science can conceive of an empirical generalization existing in some ultimate sense, but it is beyond the reach of humans. The researcher cannot declare the effect to exist or not exist. The researcher is in a position only to declare that the evidence crossed the threshold so that an effect can be said to have been found, or that the evidence did not cross the threshold and so the finding is that the hypothesis of no effect cannot be rejected.

There are two ways a researcher can be correct: the phenomenon exists and the researcher rejects the hypothesis of no effect; or the phenomenon does not exist and the researcher concludes that the hypothesis of no effect cannot be rejected. And there are two ways the researcher can be incorrect: the phenomenon exists but the researcher concludes that the hypothesis of no effect cannot be rejected; or the phenomenon does not exist but the researcher rejects the hypothesis of no effect.

The impossibility of declaring "no difference" or "no effect" stems from the fact that in making decisions under uncertainty, one must adopt a no-information starting point and evaluate the extent to which the evidence moves the decision away from that starting point and along the scale toward rejection of the hypothesis of no difference and the finding of a difference (or an effect) attributable to the hypothesized cause. In scientific hypothesis testing, that starting point is the null hypothesis.

The null hypothesis states the assumption that no differences exist between two conditions. For example, in an experiment where the independent variable is a drug administered or not administered to two groups of people (termed the experimental and control groups), and the dependent variable is how often members of each group catch colds, the null hypothesis states, for example, the assumption that there will be no difference in the mean number of colds in the experimental and control groups. Expressed as an equation: $H_0: M_E = M_C$. That is an assumption, a temporary stance, not a finding, not a conclusion. If insufficient evidence of an effect of an independent variable on a dependent variable is obtained, researchers conclude that they are unable to reject the null hypothesis. But they cannot embrace the null as an affirmative conclusion.\footnote{Even if the data strongly suggest no effect of an independent variable on a dependent variable, in reality an effect might still exist. There are many ways to do a study poorly so that an effect that actually exists goes undetected. These include insufficiently sensitive measures, a sample size that is too small, statistical tests that have too little power to detect the effect, confounding that suppresses the relationship between the independent and dependent variables of interest, and numerous other errors more subtle as well as more blatant. In short, the researcher cannot know everything there might be to know about a phenomenon. The possibility of an effect always exists. All the researcher can assess is whether or not the data in the study at hand have crossed the threshold required for rejection of the null hypothesis. As a formal matter, the null hypothesis can never be proven. As a practical matter, however, if enough well-designed and well-conducted studies of the same phenomenon are conducted, researchers will come to believe that $X$ does not cause $Y$. Similarly, consider KARL POPPER, CONJECTURES AND REPUTATIONS: THE GROWTH OF SCIENTIFIC KNOWLEDGE (5th ed.) (1989) (arguing that theories can never be proven true, but can only be disproved, though a theory that has survived many sound attempts at disproof comes to be regarded as valid).} Put succinctly, science recognizes that failing to reject the null hypothesis is not the same as accepting the null as true.

Another decision tool is needed: a threshold beyond which the evidence changes the conclusion from non-rejection of the null to a rejection of the null hypothesis. For science using statistical hypothesis testing, that is the \textit{significance level} (also known as the alpha level or p-
level). Researchers are to conclude that a relationship between an independent and dependent variable exists if and only if the evidence crosses the line defined by that significance level. A typical significance level is \( p < .05 \), meaning that the average difference between (for example) the dependent variable measured in the experimental and control groups is such that it will occur by chance less of than 5% of the time. That is, if the observed difference is due to random variation rather than a real effect of the independent variable, we would observe the difference in fewer than five out of every hundred samples. Because scientists typically quantify their observations, they can use those measures to calculate the risk of a false rejection of the null hypothesis, and that risk of error (in the example here, and usually) is kept below 5%.

Science is not indifferent to the two kinds of errors mentioned above. It recognizes that the two kinds of error stand in some ratio to one another. Erroneous rejections of the null hypothesis would lead to conclusions that phenomena exist when they do not actually exist, and theories would be constructed or renovated to incorporate those findings. That would lead to a great deal of error and confusion, so false rejection of the null is anathema to science, and the risk of such errors is kept low (below 5%). Science is inherently conservative in the sense that it would prefer to overlook phenomena which actually do exist (non-rejection of a false null) than to recognize as true phenomena which in actuality do not exist (rejection of a true null).

Scientists and statisticians realize that where they set the threshold for rejecting the null hypothesis changes the risk of one type of error relative to the other type of error.

The description in the preceding paragraph pertains to pure science, the enterprise of knowledge-building for its own sake. In that endeavor, avoidance of false rejections of the null is understandably favored. But consider an instance of applied science, for example where one is testing a cure for a dreaded disease. Here it would make sense to run a greater risk of erroneously concluding that a treatment works when in reality it does not (effectuated by setting the \( p \)-level at, perhaps, \( p < .10 \) or \( p < .20 \)), than to erroneously conclude that a treatment does not work when in reality it is effective. Nevertheless, in practice, researchers of all kinds usually adhere to the conventional \( p \)-levels of 5% or 1%.  

C. Comparing Legal and Scientific Decision-making Under Uncertainty

To make our point most clearly, in Figure 3 we place the two matrices side by side and briefly describe the parallel structure of decision-making under uncertainty in legal trials and in scientific hypothesis testing.

To begin, the literal structure of the matrices is the same. First, each represents a true state of reality which can never be known directly (that a defendant is innocent or guilty; that a null hypothesis is true or false). Second, each involves a decision that has to be made on the basis of incomplete and imperfect evidence (whether a defendant should be found guilty or not guilty; whether a null hypothesis should be rejected or not rejected).

---

8 Thus, the law appears to be more sophisticated than science in its development of different decision thresholds – preponderance, clear and convincing, beyond a reasonable doubt – for decisions involving consequences with different utilities and disutilities.
Each recognizes two types of error. In law: erroneous conviction and erroneous acquittal. In science: erroneous rejection of the null hypothesis and erroneous failure to reject the null.

Each recognizes that one of those types of error has conspicuously greater disutility than the other, and the decision process is constructed around avoidance of that more salient error – so much so that the other type of error almost recedes from view. In law, wrongful convictions are regarded as the more regrettable and less tolerable of the two types of error. In science, erroneous rejection of the null hypothesis (also termed "Type I error") is the more serious and less tolerable type of error. Erroneous failure to reject the null hypothesis ("Type II error") is the parallel of erroneous acquittal – unfortunate and undesirable, but not nearly so regrettable as the opposite type of error.

Both law and science begin their decision processes by defining a no-information starting point. For law, that is the presumption of non-liability (in criminal law: the presumption of innocence). For science that is the null hypothesis.

Both law and science set a threshold which, if crossed, the starting point is to be rejected. For the law, that is the standard of proof. In criminal law, it is the threshold of reasonable doubt. The risk of erroneous conviction is reduced by setting the threshold of reasonable doubt high. For science, the threshold is the significance level (also known as the p-level or alpha level). The risk of erroneous rejection of the null hypothesis is reduced by setting a fairly extreme significance level.

Both law and science clearly recognize that failing to reject the starting point is not the equivalent of accepting the starting point as true. Law says "not guilty" rather than "innocent." Science conveys the same idea with phrases like "non-significant," or that the researcher "did not reject the null hypothesis" – rather than concluding that the correct finding is that no difference exists between experimental conditions or that there is no effect of the independent variable on the dependent variable.

Because scientists of most kinds measure quantitatively what they are observing, it is possible to calculate the probability of Type I error and to compare that value to the threshold of significance, leading to a straightforward determination of whether the risk of erroneous rejection of the null hypothesis is extreme enough to reject the null. Scientists and statisticians routinely perform such calculations on their research data. Factfinders in law are expected to do the same thing (e.g., compare their certainty of guilt against the threshold of reasonable doubt), but they must do it subjectively. The fact of quantification versus subjective judgment can be regarded as a minor difference set against the background of so much important conceptual similarity.

---

9 See supra note 6.
10 Indeed, both the threshold of reasonable doubt and the most common alpha level reflect approximately the same risk of the more serious of the two types of error: something in the vicinity of 5%.
2. HISTORICAL EVOLUTION OF THE TWO MODELS

We could stop at this point and marvel at the stunning similarities in two fields usually regarded as so different. But it is hard to resist the temptation to inquire into the origins of these two approaches to decision-making under uncertainty which are really the same approach. Did law and science inform each other in any way, or did their models evolve separately? Because the legal model is older, if either model were based upon the other, it would have to be that the scientists and statisticians who developed the scientific model borrowed from the legal model. Looking into the history of their origins provides clues to the answer.

A. Origins of the Presumption of Innocence and the Reasonable Doubt Standard of Proof

In law, the presumption of innocence has ancient origins. Langbein has summarized: The presumption of innocence "was known from classical Roman Law and had been reinvigorated in the natural law literature of the Seventeenth Century. English juristic writers subscribed to it, from Fortescue to Coke to Blackstone."11

The reasonable doubt standard of proof developed later. Although glimmers of the notion appear inconsistently in cases going back centuries,12 the rule did not crystallize in Anglo-American law until the last quarter of the Eighteenth Century.13 The reasonable doubt standard appears to have become a familiar concept in English trials from the mid-1780s, though it took somewhat longer to become established as a rule of law.14

Whether the beyond-a-reasonable-doubt standard of proof in law emerged in the late Eighteenth Century or sooner than that, it clearly preceded the emergence of parallel concepts in science and statistics, and therefore could not have derived from the non-legal uses. The social needs which gave rise to the reasonable doubt standard, as understood by recent historical scholarship, are worth examining.15

The reasonable doubt standard appears to have originated not to protect criminal defendants from erroneous conviction, but as one of the numerous responsibility-shifting and agency-denying devices adopted by the law to benefit the decision-makers, insulating them from moral responsibility so that they would be less fearful of making the decisions they were called

---

12 Jurists in criminal trials sometimes made such pronouncements as that the evidence must be "clearer than the noonday sun" to convict a defendant (that is, admitting of no, or virtually no, doubt). E.g., Julius Clarus (1585) ("Debent autem esse in criminalibus probationes luce meridiana clariores... [e]t hoc omnes sciunt et dicunt"). See also, Anthony A. Morano, A Reexamination of the Development of the Reasonable Doubt Rule, 55 BOSTON UNIV. L. REV. 507 (1975), finding reasonable doubt employed as early as 1770 in Massachusetts, in the Boston Massacre trials.
13 Langbein, supra note 11, at 262 ("In routine criminal adjudication at the Old Bailey we find scant indication of any standard of proof being the subject of jury instruction until the last quarter of the eighteenth century.")
14 That is, that it moved from application according to the discretion of individual judges to being invariably applied by all judges.
upon to deliver. More specifically, the reasonable-doubt standard developed as a device for giving jurors what Whitman terms the "moral comfort" to convict.

Until the early Thirteenth Century, judges in Christian Europe were fearful about performing their role because doing so violated New Testament injunctions against judging and of causing bloodshed. Christians of that time and place believed that taking any part in causing the death of an accused, or inflicting other blood punishments, however justifiable, might lead to eternal damnation. Bloodshed was viewed as a moral pollution that barred those responsible from receiving communion for three years and necessitated their purification. Even soldiers who killed in legitimate defense were considered to have become morally polluted.

For centuries, the solution for judges was the trial by ordeal. By subjecting defendants to ordeals – for example throwing them into water to see if they sank (indicating innocence) or floated (indicating guilt) – judges could avoid responsibility and moral consequences of deciding cases by invoking God's judgment instead of their own. Such divine invocation, however, required the participation of the clergy. This passed the moral jeopardy from judges to clerics, who feared that if the procedure were not correct, and the ordeal led to a verdict that did not accurately reflect God's judgment, then it was the priests who would become tainted. The Church's fears about the moral pollution of its priests led the Fourth Lateran Council of the Christian Church to safeguard the purity of the clergy by adopting Canon 18, which prohibited priests from participating in judicial ordeals.

What were judges and courts to do now? New solutions were needed to elude the moral consequences of decision. In Continental Europe, Canon 18 led to the development of inquisitorial judicial procedures. By "inquiring" of persons suspected of crimes, the process did not require the participation of accusers or witnesses. So long as the judge did not supply any of the evidence, or rely on any personal knowledge, judges were considered to be maintaining a safe moral distance from the bloody consequences of judicial decisions. The defendant's confession led to the verdict and caused the punishment. To maintain proper moral distance, judges needed to follow procedural rules rigidly, and canon lawyers developed the rule "in dubio pro reo" (in doubt you must decide for the defendant). This rule "created a form of protection for the accused that grew out of the familiar fear that the judge might make himself into a murderer." In England, the departure of the institution of the ordeal led to the invention of the jury – first consisting of men from the community with knowledge of the facts at issue (self-informing juries), and eventually to juries constituted and operated in ways more familiar to us today. Passing the moral risks of decision-making to jurors solved the judge's problem by creating problems for the jurors. Various rules came into being to help ensure that jurors would, when appropriate, convict defendants of serious crimes. These included fines and imprisonment for

16 Found in the Gospel of Matthew.
17 Whitman, supra note 15, at chapter 2.
18 Id., at 122.
19 Id., at 133.
20 "Eventually" being the Nineteenth Century.
jurors who refused to render a verdict, special verdicts (facts found but no judgments made), immunity from penalties for erroneous convictions, allowance of verdicts other than blood punishments, and the development of new punishments that avoided bloodshed. From 1718 until the outbreak of the American Revolution, transportation to the American colonies was the most common punishment imposed by English courts.

By rendering transportation impossible, the American Revolution created a serious problem for English courts, whose juries again faced the moral risks of being asked to reach verdicts that would lead to blood punishments. The solution was the development of the reasonable doubt standard of proof, "a formula intended to ease the fears of jurors who might otherwise refuse to pronounce the defendant guilty." If proper procedures were followed, and if there were no reasonable doubt as to the accused's guilt, then jurors could be comforted that the defendant was convicted by the law and the facts, and not by the jurors themselves. Thus, although today the law views the standard of proof as a device to protect defendants from wrongful conviction, the historical development suggests that its real purpose was to protect jurors from other-worldly consequences of their decisions.

Though the law's version of these concepts was in place by the end of the Eighteenth Century, science and statistics had to wait until the curtain rose on the Twentieth Century.

B. Origins of Statistical Significance Testing and the Null Hypothesis

Science researchers have employed statistical analyses since at least the middle of the Eighteenth Century. Among the first scientists to employ statistics were astronomers, who used them to determine the coefficients in the equations for the movement of planets. In 1835, the French physician Pierre Louis used statistics to show that bleeding patients was not as effective a way to treat pneumonia as was then believed, and in 1855, the English physician John Snow analyzed a natural experiment to demonstrate that cholera is likely an infectious disease, rather than a result of the body’s unbalanced humors. In 1860, the psychophysicist Gustav Fechner introduced statistics to the field of psychology in his experiments on sensation and perceptual thresholds of weight detection. Despite a relatively long history in the sciences, however, statistical analyses did not much resemble the current model for hypothesis testing until the early Twentieth Century.

In 1908, a groundbreaking statistical paper emerged from an unlikely enterprise. In breweries all across Europe, despite precise techniques and lengthy apprenticeships, beer-making was still a hit-or-miss process. At Ireland’s Guinness Brewery, though, Cecil Guinness was becoming the first beer-maker to appreciate the value of scientific experimentation in the pursuit of a consistent and superior brew. Guinness implemented the forward-thinking policy of finding

21 Id., at 193.
24 Stigler, supra note 22.
the best and brightest chemistry graduates from Oxford and Cambridge, and hiring them as brewers. In 1899, William Sealy Gosset, a recent Oxford graduate, was brought on as one of these recruits.

Gosset and his colleagues began experimenting on the effects of a variety of factors – hops, barleycorn size, amount of rainfall, temperature, and so on – on the quality of the beer, and soon realized that they could not use their results to make conclusive statements. The problem was that their experiments did not have enough data points to use the standard statistical analyses of the time. Statisticians in the early 1900s used formulas that assumed the standard deviation of the sample was the same as the true standard deviation of the population, which was not problematic when the sample sizes were large enough. But Gosset realized that this assumption was not valid for his small experimental sample sizes. After consultation with Karl Pearson, a leading biometrician, Gosset developed a statistical analysis that took into account the size of the samples and allowed valid comparison of two smaller groups. Guinness did not want Gosset to publish the paper under his own name, apparently fearing that doing so would reveal that the secret behind Guinness’s famous brew was scientific testing and statistics. Therefore, Gosset had the paper published under the name “Student,” and his influential and still widely used $t$-test entered the world as “Student’s $t$-test.”

In 1912, while Gosset continued his work in brewing and statistics, a brilliant Cambridge mathematics undergraduate named Ronald A. Fisher was developing his own statistical formula. This formula, which later became known as the method of maximum likelihood, used a different calculation for standard deviation than the one Gosset had used. Fisher contacted Gosset to discuss the discrepancy, and used mathematical proofs to show that his own formula was the correct one. This started Fisher down a highly prolific road of revolutionary mathematical research and proofs, and in 1925, he published his notable book, *Statistical Methods for Research Workers*. The book was intended, as the name implies, for researchers; so, while the formulas throughout the book were highly complex, readers did not need to understand the mathematics involved to learn how to use the tools they provided.

In his book, Fisher extended application of Student’s $t$-test to testing regression coefficients and analysis of variance, and he also advocated a standard criterion for determining whether a test was significant or not. When testing a null hypothesis, researchers run statistical analyses that yield a $p$-value, which is the probability that the data they gathered would have been found in a population in which the null hypothesis was true.

Previously, researchers would look at their $p$-values and determine, somewhat arbitrarily, whether they were “small enough.” Fisher recommended using a standard 0.05 cutoff point, to be known as the “alpha level;” a researcher could set his alpha level at 0.01, instead, if he wanted his test to be very stringent. Fisher’s book was a great success, and the use of his statistical formulas became commonplace in many areas of research, particularly the social, behavioral, and biological sciences.

---

Not long after the publication of Fisher’s book, the statisticians Jerzy Neyman and Egon Pearson (son of the aforementioned biometrician Karl Pearson) wondered if there were ways to build on Fisher’s methods. They were impressed with Fisher’s idea of the standard \( p \)-value cutoffs and, furthermore, they realized that there was a second side to the usual statistical story of accepting or rejecting a single hypothesis: If a hypothesis is being tested, and the decision is made to reject it, then theoretically an alternative hypothesis is being supported. These competing hypotheses became known as the “null hypothesis,” the hypothesis of no difference, and the “alternative hypothesis,” which states that the two groups are different from each other. Although the alternative hypothesis is the hypothesis of interest, the null hypothesis is the one that is tested. Thus, hypothesis testing is indirect.

In two influential papers, Neyman and Pearson\(^27\) extended this line of reasoning, arguing that in hypothesis testing, two types of error are possible. The first, which they named “Type I error,” is rejection of the null hypothesis when the two groups are in fact the same. This is also known as a “false positive,” because it leads researchers to conclude that their alternative hypothesis is supported when it is not really true. The second, “Type II error,” is failure to reject the null hypothesis when the two groups are in fact different from each other. This is also known as a “false negative” because a researcher making this error would reject the alternative hypothesis though it actually is true.

Neyman and Pearson acknowledged that statistics would not allow researchers to ascertain the answer as it truly is in the population, but, “without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not too often be wrong.”\(^28\) They argued that the best hypothesis test would be one that balanced the two types of errors at an acceptable level, and they concluded that Fisher’s alpha level of 0.05 was appropriate. This general framework for hypothesis testing is stunningly similar to the one that had been in place in the legal world for more than a century, and remains the predominant type of hypothesis testing used today.

Of interest to the present paper is whether any of the scientists or statisticians working on this model of hypothesis testing were aware that a very similar model already existed in law. Some statistical papers make reference to the law as an illustration of decision-making under uncertain circumstances. Neyman and Pearson, for example, refer to the mathematician Laplace’s question regarding judge and jury decision-making, and briefly discuss the implications of errors of sending innocent people to prison versus setting guilty people free.\(^29\) However, we could find no discussion of the precise and nuanced legal conceptions that might have led to the similar conceptions adopted by the statisticians. The statistical literature we reviewed contained no reference to legal arguments, no citation of judicial opinions or legal commentary or legal professionals, and no use of legal terminology. This suggests that the


\(^{28}\) \textit{Id.}, \textit{On the Use the Interpretation of Certain Test Criteria for Purposes of Statistical Inference}, at 291.

\(^{29}\) \textit{Id.}, at 296.
statisticians had only a lay understanding of the model employed in legal proceedings, and were unaware of the strong similarity between the model being developed in statistics and the one already in place in law.

III. CONCLUSION

Our search for indications of any shared ancestry in the development of the legal and the scientific models of decision-making under uncertainty turned up no evidence of cross-fertilization. We found no indications that members of either field even considered the parallel problems each faced in their decision-making models at the time of their conception, nor is there much indication that even today more than a handful of members of either field are aware of the remarkable similarity in their solutions. If scientific decision-making was not informed by the legal decision-making model that had become crystallized more than a century earlier, then two virtually identical models arose independently in wholly unrelated fields. This in itself is noteworthy, and represents an unappreciated commonality between law and science.

A. Biological Evolution and Cultural Evolution

We might analogize what has happened to something that occurs in biological evolution. The development of similar adaptations in two unrelated species is termed convergent evolution. Such evolution occurs when these unrelated species experience the same kinds of selection pressures, which are then met by the each species in similar ways. A familiar example is convergent evolution in birds and bats. Both have wings, even though their most recent common ancestors did not, indicating that both species separately responded to their environments by developing this adaptation. The specific characteristics of their wings differ – for example, bird wings have feathers and bat wings have fur – but the function of wings is the same for both species. The biological world presents many instances of such similarity of response to similarity of environment.

Perhaps this concept from biological evolution can be borrowed to characterize something that happens in the "evolution" of intellectual inventions, and suggest that where the same underlying problem is encountered in different subject matter areas, similar solutions tend to be found. Though the solutions might appear different on the surface, a deeper look can discover that their structures as well as their functions are essentially the same.

30 Both parallel and convergent evolution refer to the result of evolutionary processes that brought unrelated species to have similar features. If their respective ancestors had similar features which evolved separately so that their descendants had new, different similar features, the process is termed parallel evolution (same is transformed into same). But if their respective ancestors had different features which evolved separately so that their descendants came to have new features that were similar, the process is termed convergent evolution (different is transformed into same). See, DOUGLAS J. FUTUYMA, EVOLUTIONARY BIOLOGY (3rd ed.) (1997). Convergent evolution seems a better fit to the analogous process that has occurred in the cultural evolution of law and science, considering that the purposes of the two started out being so different, before eventually arriving at the same structure for the same purpose.
B. Differences Amid the Similarities

Nothing we have written should be taken to suggest that the two systems do not also possess differences, or that they are not employed differently, or that they will not evolve beyond their present forms in different directions. After all, bats and birds, aloes and agaves, and so on, are not identical.

Scientists follow the general procedures of hypothesis testing quite rigorously. That is to say, they carefully perform significance tests in the great multitude of their research. The comparable decision-making system in legal trials is more of a heuristic, a general model that is used by judges and imposed on jurors, with limited elaboration through instructions, and the details of actually carrying out the "test" left largely unspecified.

Statisticians have a large collection of different statistical significance tests adapted to various different kinds of data and employed in specified circumstances. Judges and jurors, by contrast, do whatever they will or can with the varied kinds of evidence marshaled by the parties, while remaining within the general framework that has been the subject of this article.

The kinds of evidence used by statisticians is overwhelmingly quantitative – measures or counts of similar things, usually sampled in systematic, representative fashion from the population of similar things being studied. By contrast, the kinds of evidence brought to court are of a varied and messy lot: witnesses, documents, exhibits, expert opinions, and more. Furthermore, the evidence gathered by lawyers is anything but representative of the complete "population" of evidence relevant to a case. The adversary process expects, almost demands, that the parties seek out the most favorable evidence they can find supporting their respective positions. Thus, while scientists usually bring to statistical hypothesis testing data which form normal, bell-shaped, balanced distributions; to the trial process lawyers usually offer up bimodal distributions.

Despite such differences in the details, the similarities remain striking.

C. New Approaches to Decision-making Under Uncertainty?

Some might argue that the worlds of trials and of scientific hypothesis testing have moved beyond the tools we have described in this paper. They might point in particular to applications of Bayes' Theorem in both worlds, seen by Bayes' proponents as far more nuanced and adaptable and (in some sense) accurate.31

But such claims for Bayes' Theorem exaggerate the role it has come to play in both scientific research and legal trials. The simple fact is that Bayesian approaches to analyzing data are few and far between in reports of scientific research, while hypothesis testing following the

general model we have described remains by far the dominant tool.\textsuperscript{32}

Similarly, in legal trials, while such suggestions have been discussed by scholars for
decades, and Bayesianism is one of the centerpieces of “the new evidence scholarship,”\textsuperscript{33} few if
any practical applications of Bayesian approaches have come to displace the conventional
understanding of how the law conceives of factfinders reaching their verdicts.

The work that factfinders do in the course of reaching their verdicts (within that larger
traditional decision-making framework) could, it has been argued by numerous scholars, be
described and guided by Bayesian analysis.\textsuperscript{34} But psychological research on jury decision-
making has found this approach to be a poor description of what jurors actually do and cannot
predict the results they reach.\textsuperscript{35} Nor is it reflected in anything that jurors are expected to do or
advised to do by courts.\textsuperscript{36}

In short, neither in law nor in science has Bayesian thinking replaced traditional decision-
theoretic concepts for making decisions under uncertainty. Had that occurred, we would be
pointing out that this example of cultural evolution had continued further along parallel lines, or
that it had diverged, as the case might be.

\textit{D. Biological Evolution versus Cultural Evolution}

Analogies have their limits. Clearly, the developers of these legal and scientific models
had advantages over biological evolution. Biological evolution depends upon random mutation
and is a relatively slow process, whereas the development of trial procedures in law was a matter
of social problem-solving embedded in a larger culture, and the developments in statistical
theory by Fisher, Neyman, Pearson and others involved forethought, design, and revision.

Though not the products of randomness-and-selection, human rationality and creativity
are inevitably bounded, and so the respective models might be flawed. The fact that the models

\textsuperscript{32} The best support for the textual statement would be found by looking through current scientific journals or the
syllabi of basic statistics courses, or statistics textbooks. There one will find massive attention to the kind of
hypothesis testing discussed in this article, and little if any use of or discussion of Bayesian alternatives.

\textsuperscript{33} See supra note 31; Roger Park & Michael Saks, Evidence Scholarship Reconsidered: Results of the
Interdisciplinary Turn, 47 BOSTON COLLEGE L. REV. 949 (2006).

\textsuperscript{34} That is, factfinders each begin at some starting point (presumably something close to the presumption of
innocence) and update their estimates of the probability of guilt as each new piece of evidence is provided. (Or a
similar process in the civil context.)

\textsuperscript{35} See review in Joseph Sanders et al., Trial Factfinders and Expert Evidence, in MODERN SCIENTIFIC EVIDENCE
(Faigman et al., eds) (2009).

\textsuperscript{36} We want to acknowledge, however, the great value of much Bayesian thinking, and point to the role that it can
usefully play in the law. As one of the authors has noted elsewhere, the best uses of Bayesianism have arisen outside
of trial decision-making, in analyzing the benefits and costs of existing and proposed rules. Park & Saks, supra note
32, at 991-992. Or for conceptualizing rules. See, e.g., Richard O. Lempert, Modeling Relevance, 75 MICHIGAN L.
REV. 1021 (1977). Moreover, some fields of expertise use Bayes’ Theorem to analyze the evidence to which their
expertise is directed, and the Bayesian analysis is part of the foundation for their opinions, such as in analysis of the
meaning of DNA testing in parentage disputes. See, Parentage Testing, in Faigman et al., MODERN
SCIENTIFIC EVIDENCE (2009).
are so similar does not speak to their inherent suitability or adaptiveness. That both have survived for long periods of time in their respective fields does suggest that they function well to meet the problems they were developed to solve and continue to solve. But in the future they might evolve further – and diverge from each other – in order to further improve their function in their respective fields.

But, for the present, the striking similarity of law and science in their respective models of making decisions under uncertainty is worth marveling at.
Figure 1
The Structure of Decision-making Under Uncertainty in the Law

<table>
<thead>
<tr>
<th></th>
<th>Unknown State of Reality</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Guilty</td>
<td>Innocent</td>
</tr>
<tr>
<td>Decision</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>G</strong></td>
<td>Correct Conviction</td>
<td>Erro...</td>
</tr>
<tr>
<td><strong>NG</strong></td>
<td>Erroneous Acquittal</td>
<td>Correct Acquittal</td>
</tr>
</tbody>
</table>
Figure 2
The Structure of Decision-making Under Uncertainty in Science

<table>
<thead>
<tr>
<th>Decision</th>
<th>Unknown State of Reality</th>
<th>H₀ is False</th>
<th>H₀ is True</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject H₀</td>
<td>Correct Rejection</td>
<td></td>
<td>Error Type I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p&lt;0.05</td>
</tr>
<tr>
<td>Don’t Reject H₀</td>
<td>Error Type II</td>
<td>Correct Non-rejection</td>
<td></td>
</tr>
</tbody>
</table>
Figure 3
Comparing Legal and Scientific Decision-making Under Uncertainty

<table>
<thead>
<tr>
<th>Unknown State of Reality</th>
<th>Guilty</th>
<th>Innocent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct Conviction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erroneous Conviction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erroneous Acquittal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct Acquittal</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unknown State of Reality</th>
<th>H₀ is False</th>
<th>H₀ is True</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reject H₀</td>
<td>Correct Rejection</td>
<td>Error Type I</td>
</tr>
<tr>
<td>p&lt;.05</td>
<td></td>
<td>p&lt;.05</td>
</tr>
<tr>
<td>Don’t Reject H₀</td>
<td>Error Type II</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Correct Non-rejection</td>
<td></td>
</tr>
</tbody>
</table>

Same structure
Two types of error
Recognition that one type of error is more serious than the other
No-information starting point (presumption of non-liability = null hypothesis)
Specified threshold for rejection of starting point (standard of proof = p-level)
Different thresholds reflect different disutilities for different errors
Reject starting point only when sufficient evidence exists
Failing to reject starting point is not equivalent to accepting the starting point