# Testing by betting: A strategy for statistical and scientific communication

Glenn Shafer

*Rutgers University, Newark, New Jersey, USA*

E-mail: gshafer@business.rutgers.edu

[To be read before The Royal Statistical Society at the Society's 2020 annual conference held online on Wednesday, September 9th, 2020, the President, Professor D. Ashby, in the Chair]

**Summary**. The most widely used concept of statistical inference — the p-value — is too complicated for effective communication to a wide audience. This paper introduces a simpler way of reporting statistical evidence: report the outcome of a bet against the null hypothesis. This leads to a new role for likelihood, to alternatives to power and confidence, and to a framework for meta-analysis that accommodates both planned and opportunistic testing of statistical hypotheses and probabilistic forecasts. This framework builds on the foundation for mathematical probability developed in previous work by Vladimir Vovk and myself.

Keywords: betting score, game-theoretic probability, likelihood ratio, p-value, statistical communication, warranty

## 1. Introduction

The most widely used concept of statistical inference — the p-value — is too complicated for effective communication to a wide audience (McShane and Gal, 2017; Gigerenzer, 2018). This paper introduces a simpler way of reporting statistical evidence: report the outcome of a bet against the null hypothesis. This leads to a new role for likelihood, to alternatives to power and confidence, and to a framework for meta-analysis that accommodates both planned and opportunistic testing of statistical hypotheses and probabilistic forecasts.

Testing a hypothesized probability distribution by betting is straightforward. We select a nonnegative payoff and buy it for its hypothesized expected value. If this bet multiplies the money it risks by a large factor, we have evidence against the hypothesis, and the factor measures the strength of this evidence. Multiplying our money by 5 might merit attention; multiplying it by 100 or by 1000 might be considered conclusive.

The factor by which we multiply the money we risk — we may call it the *betting score* — is conceptually simpler than a p-value, because it reports the result of a single bet, whereas a p-value is based on a family of tests. As explained in Section 2, betting scores also have a number of other advantages:

(a) Whereas the certainty provided by a p-value is sometimes exaggerated, the uncertainty remaining when a large betting score is obtained is less easily minimized. Whether or not you have been schooled in mathematical statistics, you will not forget that a long shot can succeed by sheer luck.
(b) A bet (a payoff selected and bought) determines an implied alternative hypothesis, and the betting score is the likelihood ratio with respect to this alternative. So the evidential meaning of betting scores is aligned with our intuitions about likelihood ratios.
(c) Along with its implied alternative hypothesis, a bet determines an implied target: a value for the betting score that can be hoped for under the alternative hypothesis. Implied targets can be more useful than power calculations, because an implied target along with an actual betting score tells a coherent story. The notion of power, because it requires a fixed significance level, does not similarly cohere with the notion of a p-value.
(d) Testing by betting permits opportunistic searches for significance, because the persuasiveness of having multiplied one's money by successive bets does not depend on having followed a complete betting strategy laid out in advance.

A shift from reporting p-values to reporting outcomes of bets cannot happen overnight, and the notion of calculating a p-value will always be on the table when statisticians look at the standard or probable error of the estimate of a difference; this was already true in the 1830s (Shafer, 2019). We will want, therefore, to relate the scale for measuring evidence provided by a p-value to the scale provided by a betting score. Any rule for translating from the one scale to the other will be arbitrary, but it may nevertheless be useful to establish some such rule as a standard. This issue is discussed in Section 3.

Section 4 considers statistical modeling and estimation. A statistical model encodes partial knowledge of a probability distribution. In the corresponding betting story, the statistician has partial information about what is happening in a betting game. We see outcomes, but we do not see what bets have been offered on them and which of these bets have been taken up. We can nevertheless equate the model's validity with the futility of betting against it. A strategy for a hypothetical bettor inside the game, together with the outcomes we see, then translates into *warranties* about the validity of the bets that were offered. The strategy tells the bettor what bets to make as a function of those offered, and if the game involves the bettor's being offered any payoff at the price given by a probability distribution — the distribution remaining unknown to us, because we are not inside the game — then assertions about the validity of this unknown probability distribution are warrantied. Instead of $(1-\alpha)$-confidence in an assertion about the distribution, we obtain a $(1/\alpha)$-warranty. Either the warrantied assertion holds or the hypothetical bettor has multiplied the money he risked by $1/\alpha$.

A statement of $(1-\alpha)$-confidence can be interpreted as a $(1/\alpha)$-warranty, one resulting from all-or-nothing bets. But the more general concept of warranty obtained by allowing bets that are not all-or-nothing has several advantages:

(a) Like individual betting scores, it gives colour to residual uncertainty by evoking our knowledge of gambling and its dangers.

(b) Observations together with a strategy for the bettor produce more than one warranty set. They produce a $(1/\alpha)$-warranty set for every $\alpha$, and these warranty sets are nested.

(c) Because it is always legitimate to continue betting with whatever capital remains, the hypothetical bettor can continue betting on additional outcomes, and we can update our warranty sets accordingly without being accused of "sampling to a foregone conclusion". The same principles authorize us to combine warranty sets based on successive studies.

The conclusion of the paper (Section 5) summarizes the advantages of testing by betting. An appendix (Section 6) situates the idea in the broader landscape of theoretical statistics and other proposed remedies for the misunderstanding and misuse of p-values and significance testing.

For further discussion of betting as a foundation for mathematical probability, statistics, and finance, see Shafer and Vovk (2019) and related working papers at `www.probabilityandfinance.com`. This paper draws on some of the mathematical results reported in Chapter 10 of Shafer and Vovk (2019), but the crucial concepts of implied alternative, implied target, and warranty are newly introduced here.

## 2. Testing by betting

You claim that a probability distribution $P$ describes a certain phenomenon $Y$. How can you give content to your claim, and how can I challenge it?

Assuming that we will later see $Y$'s actual value $y$, a natural way to proceed is to interpret your claim as a collection of betting offers. You offer to sell me any payoff $S(Y)$ for its expected value, $\mathbf{E}_P(S)$. I choose a nonnegative payoff $S$, so that $\mathbf{E}_P(S)$ is all I risk. Let us call $S$ my *bet*, and let us call the factor by which I multiply the money I risk,

$$\frac{S(y)}{\mathbf{E}_P(S)},$$

my *betting score*. This score does not change when $S$ is multiplied by a positive constant. I will usually assume, for simplicity, that $\mathbf{E}_P(S) = 1$ and hence that the score is simply $S(y)$.

A large betting score can count as evidence against $P$. What better evidence can I have? I have bet against $P$ and won. On the other hand, the possibility that I was merely lucky remains stubbornly in everyone's view. By using the language of betting, I have accepted the uncertainty involved in my test and made sure that everyone else is aware of it as well.

I need not risk a lot of money. I can risk as little as I like — so little that I am indifferent to losing it and to winning any amount the bet might yield. So this use of the language of betting is not a chapter in decision theory. It involves neither the evaluation of utilities nor any Bayesian reasoning. I am betting merely to make a point. But whether I use real money or play money, I must declare my bet before the outcome $y$ is revealed, in the situation in which you asserted $P$.

This section explains how testing by betting can bring greater flexibility and clarity into statistical testing. Section 2.1 explains how betting can be more opportunistic than conventional significance testing. Section 2.2 explains that a bet implies an alternative hypothesis, and that the betting score is the likelihood ratio with respect to this alternative. Section 2.3 explains how the alternative hypothesis in turn implies a target for the bet. Finally, Section 2.4 uses three simple but representative examples to show how the concepts of betting score and implied target provide a clear and consistent message about the result of a test, in contrast to the confusion that can arise when we use the concepts of p-value and power.

### 2.1. Basic advantages

The standard way of testing a probability distribution $P$ is to select a *significance level* $\alpha \in (0,1)$, usually small, and a set $E$ of possible values of $Y$ such that $P(Y \in E) = \alpha$. The event $E$ is the *rejection region*. The probability distribution $P$ is discredited (or *rejected*) if the actual value $y$ is in $E$.

Although textbooks seldom make the idea explicit, a standard test is often thought of as a bet: I pay \$1 for the payoff $\$S_E$ defined by

$$S_E(y) := \begin{cases} \frac{1}{\alpha} & \text{if } y \in E \\ 0 & \text{if } y \notin E. \end{cases} \tag{1}$$

If $E$ happens, I have multiplied the \$1 I risked by $1/\alpha$. This makes standard testing a special case of testing by betting, the special case where the bet is *all-or-nothing.* In return for \$1, I get either $\$(1/\alpha)$ or \$0.

Although statisticians are accustomed to all-or-nothing bets, there are two good reasons for generalizing beyond them. First, the betting score $S(y)$ from a more general bet is a graduated appraisal of the strength of the evidence against $P$. Second, when we allow more general bets, testing can be opportunistic.

*A betting outcome is a graduated appraisal of evidence.* A betting score $S(y)$ appraises the evidence against $P$. The larger $S(y)$, the stronger the evidence.

A p-value also appraises the evidence against $P$; the smaller the p-value, the stronger the evidence. But p-values are more complicated than betting scores; they involve a large class, ideally a continuum, of all-or-nothing tests. To obtain a p-value, we usually begin with function $T$ of $Y$, called a *test statistic.* In the ideal case, there exists for each significance level $\alpha \in (0,1)$ a number $t_\alpha$ such that

$$P\left(T \geq t_\alpha\right) = \alpha. \tag{2}$$

So we have an all-or-nothing test for each $\alpha$: reject $P$ if $T(y) \geq t_\alpha$. The *p-value*, say $\mathsf{p}(y)$, is the smallest $\alpha$ for which the test rejects:

$$\mathsf{p}(y) := \inf\{\alpha \mid T(y) \geq t_\alpha\} = P(T \geq T(y)). \tag{3}$$

The larger $T(y)$, the smaller $\mathsf{p}(y)$.

Large values of $T(y)$ are supposed to discredit $P$. The p-value $\mathsf{p}(y)$ locates the degree of discredit on a scale from zero to one. But what does the scale mean? For a mathematical statistician, this question is answered by (2) and (3). For less sophisticated users of statistics, the import of these equations can be elusive. The difficulty can be explained using the language of betting. *Had I known $y$ in advance*, I could have multiplied my money by $1/\mathsf{p}(y)$ by making an all-or-nothing bet with significance level $\mathsf{p}(y)$. But I did not know $y$ in advance, and pretending that I did would be cheating.

*Betting can be opportunistic.* The probabilistic predictions that can be associated with a scientific hypothesis usually go beyond a single comprehensive probability distribution. In some cases, a scientist may begin with a joint probability distribution $P$ for a sequence of variables $Y_1, \dots, Y_N$ and formulate a plan for successive experiments that will allow her to observe them. But the scientific enterprise is usually more opportunistic. A scientist might perform an experiment that produces $Y_1$'s value $y_1$ and then decide whether it is worthwhile to perform the further experiment that would produce $Y_2$'s value $y_2$. Perhaps no one even thought about $Y_2$ at the outset. One scientist or team tests the hypothesis using $Y_1$, and then, perhaps because the result is promising but not conclusive, some other scientist or team comes up with the idea of further testing the hypothesis with a second variable $Y_2$ from a hitherto uncontemplated new experiment or database.

Testing by betting can accommodate this opportunistic nature of scientific research. Imagine, for example, that I doubt the validity of the probability forecasts made by a particular weather forecaster. Imagine further that the forecaster decides each day, on a whim, what to forecast that day; perhaps he will give a probability distribution for the amount of rain, perhaps a probability distribution for the temperature at 10:00 a.m., etc. In spite of his unpredictability, I can try to show that he is a poor forecaster by betting against him. I start with $1, and each day I buy a random variable for the expected value he attributes to it. I take care never to risk more than I have accumulated so far, so that my overall risk never exceeds the $1 with which I began. If I have accumulated $1000 after a year or two, this will be convincing evidence against the forecaster.

Such opportunistic betting boils down to multiplying betting scores. My initial capital is 1. My first bet $S_1$ is nonnegative and has expected value 1 according to the forecaster. After it is settled, my capital is $S_1(y_1)$. Now I select a nonnegative bet $S_2$ to which the forecaster now gives expected value 1, and I use my current capital to buy a multiple of $S_2$. In other words, I pay $S_1(y_1)$ for $S_1(y_1)S_2$. After this second bet is settled, I have $S_1(y_1)S_2(y_2)$. (This argument assumes that the price of my second bet is exactly equal to my current capital after the first bet is settled. Since the only constraint is that I not risk my capital becoming negative, we might imagine other options. I could reserve some of my capital and buy a payoff that costs less, or perhaps I might be allowed to buy a payoff that costs more than my current capital if this payoff is bounded away from zero. But these ideas do not really increase my betting opportunities. When I am not allowed to risk my capital becoming negative, any bet I make can be understood as buying a nonnegative payoff that has expected value equal to my current capital.)

Multiplying betting scores may sometimes give a more reasonable evaluation of scientific exploration than other methods of combination. Consider the scientist who uses a significance level of 5% in a search for factors that might influence a phenomenon. Her initial explorations are promising, but only after 20 tries (20 slightly different chemicals in a medical study or 20 slightly different stimuli in a psychological study) does she find an effect that is significant at 5%. How seriously should we take this apparent discovery? One standard answer is that the significance level of 5% should be multiplied by 20; this is the Bonferroni adjustment. It has a betting rationale; we may suppose that the scientist has put up \$1 each time she tests a factor, thereby investing a total of \$20. She loses her \$1 on each of the first 19 tries, but she wins \$20 on her 20th try. When we recognize that she actually invested \$20, not merely \$1, we might conclude that her final betting score is 20/20, or 1. But this will be unfair if the first 19 experiments were promising, as the product of 20 betting scores that are only a little larger than 1 may be reasonably large.

In many fields, the increasing resources being devoted to the search for significant effects has led to widespread and justified skepticism about published statistical studies purporting to have discovered such effects. This is true for both experimental studies and studies based on databases. A recent replication of published experimental studies in social and cognitive psychology has shown that many of their results are not reliable (Collaboration, 2015). A recent massive study using databases from medical practice has shown that null hypotheses known to be true are rejected at a purported 5% level about 50% of the time (Madigan et al., 2014; Schuemie et al., 2018). A recent review of database studies in finance has noted that although a large number of factors affecting stock prices have been identified, few of these results seem to be believed, inasmuch as each study ignores the previous studies (Harvey, 2017). These developments confirm that we need to report individual statistical results in ways that embed them into broader research programs. Betting scores provide one tool for this undertaking, both for the individual scientist reporting on her own research and for the meta-analyst reporting on the research of a scientific community (ter Schure and Grünwald, 2019).

### 2.2. Score for a single bet = likelihood ratio

For simplicity, suppose $P$ is discrete. Then the assumption $\mathbf{E}_P(S) = 1$ can be written

$$\sum_y S(y)P(y) = 1.$$

Because $S(y)$ and $P(y)$ are nonnegative for all $y$, this tells us that the product $SP$ is a probability distribution. Write $Q$ for $SP$, and call $Q$ the alternative *implied* by the bet $S$. If we suppose further that $P(y) > 0$ for all $y$, then $S = Q/P$, and

$$S(y) = \frac{Q(y)}{P(y)}. \tag{4}$$

A betting score is a likelihood ratio.

Conversely, a likelihood ratio is a betting score. Indeed, if $Q$ is a probability distribution for $Y$, then $Q/P$ is a bet by our definition, because $Q/P \geq 0$ and

$$\sum_y \frac{Q(y)}{P(y)} P(y) = \sum_y Q(y) = 1.$$

According to the probability distribution $Q$, the expected gain from the bet $S$ is nonnegative. In other words, $\mathbf{E}_Q(S)$ is greater than 1, $S$'s price. In fact, as a referee has pointed out, $\mathbf{E}_Q(S) = \mathbf{E}_P(S^2)$ and hence

$$\mathbf{E}_Q(S) - 1 = \mathbf{E}_P(S^2) - (\mathbf{E}_P(S))^2 = \mathbf{Var}_P(S).$$

*When I have a hunch that Q is better...* We began with your claiming that $P$ describes the phenomenon $Y$ and my making a bet $S$ satisfying $S \geq 0$ and, for simplicity, $\mathbf{E}_P(S) = 1$. There are no other constraints on my choice of $S$. The choice may be guided by some hunch about what might work, or I may act on a whim. I may not have any alternative distribution $Q$ in mind. Perhaps I do not even believe that there is an alternative distribution that is valid as a description of $Y$.

Suppose, however, that I do have an alternative $Q$ in mind. I have a hunch that $Q$ is a valid description of $Y$. In this case, should I use $Q/P$ as my bet? The thought that I should is supported by Gibbs's inequality, which says that

$$\mathbf{E}_Q\left(\ln \frac{Q}{P}\right) \geq \mathbf{E}_Q\left(\ln \frac{R}{P}\right) \tag{5}$$

for any probability distribution $R$ for $Y$. Because any bet $S$ is of the form $R/P$ for some such $R$, (5) tells us that $\mathbf{E}_Q(\ln S)$ is maximized over $S$ by setting $S := Q/P$. Many readers will recognize $\mathbf{E}_Q\left(\ln(Q/P)\right)$ as the Kullback-Leibler divergence between $Q$ and $P$. In the terminology of Kullback's 1959 book (Kullback, 1959, p. 5), it is the mean information for discrimination in favor of $Q$ against $P$ per observation from $Q$.

Why should I choose $S$ to maximize $\mathbf{E}_Q(\ln S)$? Why not maximize $\mathbf{E}_Q(S)$? Or perhaps $Q(S \geq 20)$ or $Q(S \geq 1/\alpha)$ for some other significance level $\alpha$?

Maximizing $\mathbf{E}(\ln S)$ makes sense in a scientific context where we combine successive betting scores by multiplication. When $S$ is the product of many successive factors, maximizing $\mathbf{E}(\ln S)$ maximizes $S$'s rate of growth. This point was made famously and succinctly by John L. Kelly, Jr. (Kelly Jr., 1956, p. 926): "it is the logarithm which is additive in repeated bets and to which the law of large numbers applies." The idea has been used extensively in gambling theory (Breiman, 1961), information theory (Cover and Thomas, 1991), finance theory Luenberger (2014), and machine learning (Cesa-Bianchi and Lugosi, 2006). I am proposing that we put it to greater use in statistical testing. It provides a crucial link in this paper's argument.

We can use Kelly's insight even when betting is opportunistic and hence does not define alternative joint probabilities for successive outcomes. Even if the null hypothesis $P$ does provide joint probabilities for a phenomenon $(Y_1, Y_2, \ldots)$, successive opportunistic bets $S_1, S_2, \ldots$ against $P$ will not determine a joint alternative $Q$. Each bet $S_i$ will determine only an alternative $Q_i$ for $Y_i$ in light of the actual outcomes $y_1, \ldots, y_{n-1}$. A game-theoretic law of large numbers nevertheless holds with respect to the sequence $Q_1, Q_2, \ldots$: if they are valid in the betting sense (an opponent will not multiply their capital by a large factor betting against them), then the average of the $\ln S_i$ will approximate the average of the expected values assigned them by the $Q_i$ (Shafer and Vovk, 2019, Chapter 2).

Should we ever choose $S$ to maximize $\mathbf{E}_Q(S)$? Kelly devises a rather artificial story about gambling where maximizing $\mathbf{E}_Q(S)$ makes sense:

> ...suppose the gambler's wife allowed him to bet one dollar each week but not to reinvest his winnings. He should then maximize his expectation (expected value of capital) on each bet. He would bet all his available capital (one dollar) on the event yielding the highest expectation. With probability one he would get ahead of anyone dividing his money differently.

But when our purpose is to test $P$ against $Q$, it seldom makes sense to choose the $S$ by maximizing $\mathbf{E}_Q(S)$. As Kelly tells us, the event yielding the highest expectation under $Q$ is the value of $y_0$ for which $Q/P$ is greatest. Is a bet that risks everything on this single possible outcome a sensible test? If $Q(y_0)/P(y_0)$ is huge, much greater than we would need to refute $Q$, and yet $Q(y_0)$ is very small, then we would be buying a tiny chance of an unnecessarily huge betting score at the price of very likely getting a zero betting score even when the evidence against $P$ in favor of $Q$ is substantial.

Choosing $S$ to maximize $Q(S \geq 1/\alpha)$ is appropriate when the hypothesis being tested will not be tested again. It leads us to the Neyman-Pearson theory, to which we now turn.

*The Neyman-Pearson lemma.* In 1928, Jerzy Neyman and E. S. Pearson suggested that for a given significance level $\alpha$, we choose a rejection region $E$ such that $Q(y)/P(y)$ is at least as large for all $y \in E$ as for any $y \notin E$, where $Q$ is an alternative hypothesis (Neyman and Pearson, 1928). (Asking the reader's indulgence, I leave aside the difficulty that it may be impossible, especially if $P$ and $Q$ are discrete, to do this precisely or uniquely.) Let us call the bet $S_E$ with this choice of $E$ the *level-$\alpha$ Neyman-Pearson bet* against $P$ with respect to $Q$. The *Neyman-Pearson lemma* says that this choice of $E$ maximizes

$$Q(\text{test rejects } P) = Q(Y \in E) = Q(S_E(Y) \geq 1/\alpha),$$

which we call the *power* of the test with respect to $Q$. In fact, $S_E$ with this choice of $E$ maximizes $Q(S(Y) \geq 1/\alpha)$ over all bets $S$, not merely over all-or-nothing bets.

PROOF. If $S \geq 0$, $E_P(S) = 1$, $0 < \alpha < 1$, and $Q(S \geq 1/\alpha) > 0$, define an all-or-nothing bet $S'$ by

$$S'(y) := \begin{cases} \frac{1}{\alpha} & \text{if } S(y) \geq \frac{1}{\alpha} \\ 0 & \text{if } S(y) < \frac{1}{\alpha}. \end{cases}$$

Then $E_P(S') < 1$ and $Q(S' \geq 1/\alpha) = Q(S \geq 1/\alpha)$. Dividing $S'$ by $E_P(S')$, we obtain an all-nothing bet with expected value 1 under $P$ and a greater probability of exceeding $1/\alpha$ under $Q$ than $S$.

It does not maximize $\mathbf{E}_Q(\ln S)$ unless $Q = S_E P$, and this is usually an unreasonable choice for $Q$, because it gives probability one to $E$.

It follows from Markov's inequality that when the level-$\alpha$ Neyman-Pearson bet against $P$ with respect to $Q$ just barely succeeds, the bet $Q/P$ succeeds less: it multiplies the money risked by a smaller factor.

Proof. Again leaving aside complications that arise from the discreteness of the probability distributions, suppose that the rejection region $E$ for the level-$\alpha$ Neyman-Pearson test consists of all $y$ such that $Q(y)/P(y) \geq Q(y_0)/P(y_0)$, where $y_0$ is the value for which the test just barely rejects. Then, using Markov's inequality, we obtain

$$\alpha = P(E) = P\left(\frac{Q}{P} \geq \frac{Q(y_0)}{P(y_0)}\right) = P\left(\frac{Q}{P} \geq \frac{Q(y_0)}{P(y_0)} \mathbf{E}_P\left(\frac{Q}{P}\right)\right) \leq \frac{P(y_0)}{Q(y_0}.$$

But the success of the Neyman-Pearson bet may be unconvincing in such cases; see Examples 1 and 2 in Section 2.4.

R. A. Fisher famously criticized Neyman and Pearson for confusing the scientific enterprise with the problem of "making decisions in an acceptance procedure" (Fisher, 1956, Chapter 4). Going beyond all-or-nothing tests to general testing by betting is a way of taking this criticism seriously. The choice to "reject" or "accept" is imposed when we are testing a widget that is to be put on sale or returned to the factory for rework, never in either case to be tested again. But in many cases scientists are testing a hypothesis that may be tested again many times in many ways.

*When the bet loses money...* In the second paragraph of the introduction, I suggested that a betting score of 5 casts enough doubt on the hypothesis being tested to merit attention. We can elaborate on this by noting that a value of 5 or more for $S(y)$ means, according to (4), that the outcome $y$ was at least 5 times as likely under the alternative hypothesis $Q$ than under the null hypothesis $P$.

Suppose we obtain an equally extreme result in the opposite direction: $S(y)$ comes out less than $1/5$. Does this provide enough evidence in favor of $P$ to merit attention? Maybe and maybe not. A low value of $S(y)$ does suggest that $P$ describes the phenomenon better than $Q$. But $Q$ may or may not be the only plausible alternative. It is the alternative for which the bet $S$ is optimal in a certain sense. But as I have emphasized, we may have chosen $S$ blindly or on a whim, without

**Table 1.** Elements of a study that tests a probability distribution by betting. The proposed study may be considered meritorious and perhaps even publishable regardless of its outcome when the implied target is reasonably large and both the null hypothesis $P$ and the implied alternative $Q$ are initially plausible. A large betting score then discredits the null hypothesis.

| | name | notation |
|---|---|---|
| **Proposed study** | | |
| initially unknown outcome | phenomenon | $Y$ |
| probability distribution for $Y$ | null hypothesis | $P$ |
| nonnegative function of $Y$ with expected value 1 under $P$ | bet | $S$ |
| $SP$ | implied alternative | $Q$ |
| $\exp(\mathbf{E}_Q(\ln S))$ | implied target | $S^*$ |
| **Results** | | |
| actual value of $Y$ | outcome | $y$ |
| factor by which money risked has been multiplied | betting score | $S(y)$ |

any real opinion or clue as to what alternative we should consider. In this case, the message of a low betting score is not that $P$ is supported by the evidence but that we should try a rather different bet the next time we test $P$. This understanding of the matter accords with Fisher's contention that testing usually precedes the formulation of alternative hypotheses in science (Bennett, 1990, p. 246),(Senn, 2011, p. 57).

### 2.3. Implied targets

How impressive a betting score can a scientist hope to obtain with a particular bet $S$ against $P$? As we have seen, the choice of $S$ defines an alternative probability distribution, $Q = SP$, and $S$ is the bet against $P$ that maximizes $\mathbf{E}_Q(\ln S)$. If the scientist who has chosen $S$ takes $Q$ seriously, then she might hope for a betting score whose logarithm is in the ballpark of $\mathbf{E}_Q(\ln S)$ — i.e., a betting score in the ballpark of

$$S^* := \exp(\mathbf{E}_Q(\ln S)).$$

Let us call $S^*$ the *implied target* of the bet $S$. By (5), $S^*$ cannot be less than 1. The implied target of the all-or-nothing bet (1) is always $1/\alpha$, but as we have already noticed, that bet's implied $Q$ is not usually a reasonable hypothesis.

The notion of an implied target is analogous to Neyman and Pearson's notion of power with respect to a particular alternative. But it has the advantage that the scientist cannot avoid discussing it by refusing to specify a particular alternative. The implied alternative $Q$ and the implied target $S^*$ are determined as soon as the distribution $P$ and the bet $S$ are specified. The implied target can be computed without even mentioning $Q$, because

$$\mathbf{E}_Q(\ln S) = \sum_y Q(y) \ln S(y) = \sum_y P(y)S(y) \ln S(y) = \mathbf{E}_P(S \ln S).$$

If bets become a standard way of testing probability distributions, the implied target will inevitably be provided by the software that implements such tests, and referees and editors will inevitably demand that it be included in any publication of results. Even if the scientist has chosen her bet $S$ on a hunch and is not committed in any way to $Q$, it is the hypothesis under which $S$ is optimal, and a proposed test will not be interesting to others if it cannot be expected to achieve much even when it is optimal.

On the other hand, if the implied alternative is seen as reasonable and interesting in its own right, and if the implied target is high, then a proposed study may merit publication regardless of how the betting score comes out (see Table 1). In these circumstances, even a low betting score will be informative, as it suggests that the implied alternative is no better than the null. This feature of testing by betting may help mitigate the problem of publication bias.

### 2.4. Elementary examples

Aside from the search for significance — now often called "p-hacking" – these three misuses of p-values merit particular attention:

(a) An estimate is statistically and practically significant but hopelessly contaminated with noise. Andrew Gelman and John Carlin contend that this case "is central to the recent replication crisis in science" (Gelman and Carlin, 2017, p. 900).

(b) A test with a conventional significance level and high power against a very distinct alternative rejects the null hypothesis with a borderline outcome even though the likelihood ratio favours the null (Dempster, 1997, pp. 249–250).

(c) A high p-value is interpreted as evidence for the null hypothesis. Although such an interpretation is never countenanced by theoretical statisticians, it is distressingly common in some areas of application (Amrhein et al., 2019; Cready, 2019; Cready et al., 2019).

To see how betting scores can help forestall these misuses, it suffices to consider elementary examples. Here I will consider examples where the null and alternative distributions of the test statistic are normal with the same variance.

*Example 1.* Suppose $P$ says that $Y$ is normal with mean 0 and standard deviation 10, $Q$ says that $Y$ is normal with mean 1 and standard deviation 10, and we observe $y = 30$.

(a) Statistician A simply calculates a p-value: $P(Y \geq 30) \approx 0.00135$. She concludes that $P$ is strongly discredited.

(b) Statistician B uses the Neyman-Pearson test with significance level $\alpha = 0.05$, which rejects $P$ when $y > 16.5$. Its power is only about 6%. Seeing $y = 30$, it does reject $P$. Had she formulated her test as a bet, she would have multiplied the money she risked by 20.

(c) Statistician C uses the bet $S$ given by

$$S(y) := \frac{q(y)}{p(y)} = \frac{(10\sqrt{2\pi})^{(-1)} \exp(-(y-1)^2/200)}{(10\sqrt{2\pi})^{(-1)} \exp(-y^2/200)} = \exp\left(\frac{2y-1}{200}\right),$$

for which

$$\mathbf{E}_Q(\ln(S)) = \mathbf{E}_Q\left(\frac{2y-1}{200}\right) = \frac{1}{200},$$

so that the implied target is $\exp(1/200) \approx 1.005$. She does a little better than this very low target; she multiplies the money she risked by $\exp(59/200) \approx 1.34$.

The power and the implied target both told us in advance that the study was a waste of time. The betting score of 1.34 confirms that little was accomplished, while the low p-value and the Neyman-Pearson rejection of $P$ give a misleading verdict in favor of $Q$.

*Example 2.* Now the case of high power and a borderline outcome: $P$ says that $Y$ is normal with mean 0 and standard deviation 10, $Q$ says that $Y$ is normal with mean 37 and standard deviation 10, and we observe $y = 16.5$.

(a) Statistician A again calculates a p-value: $P(Y \geq 16.5) \approx 0.0495$. She concludes that $P$ is discredited.
(b) Statistician B uses the Neyman-Pearson test that rejects when $y > 16.445$. This test has significance level $\alpha = 0.05$, and its power under $Q$ is almost 98%. It rejects; Statistician B multiplies the money she risked by 20.
(c) Statistician C uses the bet $S$ given by $S(y) := q(y)/p(y)$. Calculating as in the previous example, we see that $S$'s implied target is 939 and yet the betting score is only $S(16.5) = 0.477$. Rather than multiply her money, Statistician C has lost more than half of it. She concludes that the evidence from her bet very mildly favors $P$ relative to $Q$.

Assuming that $Q$ is indeed a plausible alternative, the high power and high implied target suggest that the study is meritorious. But the low p-value and the Neyman-Pearson rejection of $P$ are misleading. The betting score points in the other direction, albeit not enough to merit attention.

*Example 3.* Now the case of a non-significant outcome: $P$ says that $Y$ is normal with mean 0 and standard deviation 10, $Q$ says that $Y$ is normal with mean 20 and standard deviation 10, and we observe $y = 5$.

(a) Statistician A calculates the p-value $P(Y \geq 5) \approx 0.3085$. As this is not very small, she concludes that the study provides no evidence about $P$.
(b) Statistician B uses the Neyman-Pearson test that rejects when $y > 16.445$. This test has significance level $\alpha = 0.05$, and its power under $Q$ is about 64%. It does not reject; Statistician B loses all the money she risked.

(c) Statistician C uses the bet $S$ given by $S(y) := q(y)/p(y)$. This time $S$'s implied target is approximately 7.39 and yet the actual betting score is only $S(5) \approx 0.368$. Statistician C again loses more than half her money. She again concludes that the evidence from her bet favors $P$ relative to $Q$ but not enough to merit attention.

In this case, the power and the implied target both suggested that the study was marginal. The Neyman-Pearson conclusion was "no evidence". The bet $S$ provides the same conclusion; the score $S(y)$ favors $P$ relative to $Q$ but too weakly to merit attention.

The underlying problem in the first two examples is the mismatch between the concept of a p-value on the one hand and the concepts of a fixed significance level and power on the other. This mismatch and the confusion it engenders disappears when we replace p-values with betting scores and power with implied target. The bet, implied target, and betting score always tell a coherent story. In Example 1, the implied target close to 1 told us that the bet would not accomplish much, and the betting score close to 1 only confirmed this. In Example 2, the high implied target told us that we had a good test of $P$ relative to $Q$, and $P$'s passing this test strongly suggests that $Q$ is not better than $P$.

The problem in Example 3 is the meagerness of the interpretation available for a middling to high p-value. The theoretical statistician correctly tells us that such a p-value should be taken as "no evidence". But a scientist who has put great effort into a study will want to believe that its result signifies something. In this case, the merit of the betting score is that it blocks any erroneous claim with a concrete message: it tells us the direction the result points and how strongly.

As the three examples illustrate, the betting language does not change substantively the conclusions that an expert mathematical statistician would draw from given evidence. But it can sometimes provide a simpler and clearer way to explain these conclusions to a wider audience.

## 3. Comparing scales

The notion of a p-value retains a whiff of betting. In a passage I will quote shortly, Fisher used the word "odds" when comparing two p-values. But obtaining a p-value $\mathsf{p}(y)$ cannot be interpreted as multiplying money risked by $1/\mathsf{p}(y)$. The logic of betting requires that a bet be laid before its outcome is observed, and we cannot make the bet (1) with $\alpha = 1/\mathsf{p}(y)$ unless we already know $y$. Pretending that we had made the bet would be cheating, and some penalty for this cheating — some sort of shrinking — is needed to make $1/\mathsf{p}(y)$ comparable to a betting score.

The inadmissibility of $1/\mathsf{p}(y)$ as a betting score is confirmed by its infinite expected value under $P$. Shrinking it to make it comparable to a betting score means shrinking it to a payoff with expected value 1. In the ideal case, $\mathsf{p}(y)$ is uniformly distributed between 0 and 1 under $P$, and there are infinitely many ways of shrinking $1/\mathsf{p}(y)$ to a payoff with expected value 1. (In the general case, $\mathsf{p}(y)$ is stochastically dominated under $P$ by the uniform distribution; so the payoff will have expected

**Table 2.** Making a p-value into a betting score

| p-value | $\frac{1}{\text{p-value}}$ | $\frac{1}{\sqrt{\text{p-value}}} - 1$ |
|---|---|---|
| 0.10 | 10 | 2.2 |
| 0.05 | 20 | 3.5 |
| 0.01 | 100 | 9.0 |
| 0.005 | 200 | 13.1 |
| 0.001 | 1,000 | 30.6 |
| 0.000001 | 1,000,000 | 999 |

value 1 or less.) No one has made a convincing case for any particular choice from this infinitude; the choice is fundamentally arbitrary (Shafer and Vovk, 2019, Section 11.5). But it would useful to make some such choice, because the use of p-values will never completely disappear, and if we also use betting scores, we will find ourselves wanting to compare the two scales.

It seems reasonable to shrink p-values in a way that is monotonic, smooth, and unbounded, and the exact way of doing this will sometimes be unimportant. My favorite, only because it is easy to remember and calculate, is

$$S(y) := \frac{1}{\sqrt{\mathsf{p}(y)}} - 1. \tag{6}$$

Table 2 applies this rule to some commonly used significance levels. If we retain the conventional 5% threshold for saying that a p-value merits attention, then this table accords with the suggestion, made in the introduction to this paper, that multiplying our money by 5 merits attention. Multiplying our money by 2 or 3, or by 1/2 or 1/3 as in Examples 2 and 3 of Section 2.4, does not meet this threshold.

If we adopt a standard rule for shrinking p-values, we will have a fuller picture of what we are doing when we use a conventional test that is proposed without any alternative hypothesis being specified. Since it determines a bet, the rule for shrinking implies an alternative hypothesis.

*Example 4.* Consider Fisher's analysis of Weldon's dice data in the first edition of his *Statistical Methods for Research Workers* (Fisher, 1925, pp. 66–69). Weldon threw 12 dice together 26,306 times and recorded, for each throw, how many dice came up 5 or 6. Using this data, Fisher tested the bias of the dice in two different ways.

(a) First, he performed a $\chi^2$ goodness-of-fit test. On none of the 26,306 throws did all 12 dice come up 5 or 6, so he pooled the outcomes 11 and 12 and performed the test with 12 categories and 11 degrees of freedom. The $\chi^2$ statistic came out 40.748, and he noted that "the actual chance in this case of $\chi^2$ exceeding 40·75 if the dice had been true is ·00003."

(b) Then he noted that in the $12 \times 26{,}306 = 315{,}672$ throws of a die there were altogether 106,602 5s and 6s. The expected number is $315{,}672/3 = 105{,}224$

> with standard error 264.9, so that the observed number exceeded expectation by 5.20 times its standard error, and "a normal deviation only exceeds 5·2 times its standard error once in 5 million times."

Why is the one p-value so much less than the other? Fisher explained:

> The reason why this last test gives so much higher odds than the test for goodness of fit, is that the latter is testing for discrepancies of any kind, such, for example, as copying errors would introduce. The actual discrepancy is almost wholly due to a single item, namely, the value of $p$, and when that point is tested separately its significance is more clearly brought out.

Here $p$ is the probability of a 5 or 6, hypothesized to be 1/3.

The transformation (6) turns the p-values 0.00003 and 1 in 5 million into betting scores (to one significant figure) 200 and 2,000, respectively. This does not add much by itself, but it brings a question to the surface. The statistician has chosen particular tests and could have chosen differently. What alternative hypotheses are implied when the tests chosen are considered as bets?

For simplicity, consider Fisher's second test and the normal approximation he used. With this approximation, the frequency

$$Y := \frac{\text{total number of 5s and 6s}}{315{,}672}$$

is normally distributed under the null hypothesis $P$, with mean 1/3 and standard deviation 0.00084. The observed value $y$ is 106,602/315,672 $\approx$ 0.3377. As Fisher noted, the deviation from 1/3, 0.0044, is 5.2 times the standard deviation. The function $\mathsf{p}(y)$ for Fisher's test is

$$\mathsf{p}(y) = 2\left(1 - \Phi\left(\frac{|y - \frac{1}{3}|}{0.00084}\right)\right),$$

where $\Phi$ is the cumulative distribution function for the standard normal distribution. The density $q$ for the alternative $Q$, obtained by multiplying $P$'s normal density $p$ by (6) is symmetric around 1/3, just as $p$ is. It has the same value at 1/3 as $p$ does, but much heavier tails. The probability of a deviation of 0.0044 or more under $Q$ is still very small, but only about 1 in a thousand instead of 1 in 5 million.

A different rule for shrinking the p-value to a betting score will of course produce a different alternative hypothesis $Q$. But a wide range of rules will give roughly the same picture.

We can obtain an alternative hypothesis in the same way for the $\chi^2$ test. Whereas the distribution of the $\chi^2$ statistic is approximately normal under the null hypothesis, the alternative will again have much heavier tails. Even if we consider this alternative vaguely defined, its existence supports Joseph Berkson's classic argument for discretion when using the test (Berkson, 1938).

## 4. Betting games as statistical models

In the preceding section, we learned how a single probability distribution can be tested by betting. In this section, we look at how this mode of testing extends to testing composite hypotheses and estimating parameters.

The extension will be obvious to anyone familiar with how standard tests are extended from point to composite hypotheses and used to form confidence sets. A composite hypothesis is rejected if each of its elements is rejected, and a $(1-\alpha)$-confidence set consists of all hypotheses not rejected at level $\alpha$. But when we test by betting, it is easy to get confused about who is doing the betting, and so a greater degree of formality is helpful. This formality can be provided by the notion of a *testing protocol*, which is studied in great detail and used as a foundation for mathematical probability by Shafer and Vovk (2019). A testing protocol may prescribe betting offers or simply tell who makes them. It also tells who decides what offers to accept and who decides the outcomes. It is then the protocol, not a probability distribution or a parametric class of probability distributions, that represents the phenomenon.

According to Fisher (1922), the theory of statistical estimation begins with the assumption that the statistician has only partial knowledge of a probability distribution describing a phenomenon. She knows only that this probability distribution is in a known class $(P_\theta)_{\theta\in\Theta}$. The corresponding assumption in the betting picture is that the statistician stands outside a testing protocol, seeing only some of the moves. The parameter $\theta$ is one of the moves she does not see. The player who bets, whom we call Sceptic, does see $\theta$. A strategy for Sceptic tells him how to move depending on the value of $\theta$. The statistician can specify a strategy for Sceptic and tell him to play it. If she believes that the protocol is a valid description of the phenomenon and has no reason to think the strategy has been exceptionally lucky, she can rely on the presumption that it will not multiply the capital it risks by a large factor to claim *warranties* that resemble the direct probability statements made by 19th-century statisticians (Shafer, 2019) and confidence intervals as defined by Jerzy Neyman in the 1930s (Neyman, 1937).

When the testing protocol prescribes that Skeptic be offered bets priced by a given probability distribution $P$ or $P_\theta$, Sceptic has only one opponent — the player who decides the outcome $y$ or outcomes $y_1, y_2, \ldots$ that the statistician sees. We call that player Reality. But we can generalize the picture by introducing a player called Forecaster, who announces probabilities or more limited betting offers as play proceeds. This generalization, studied at length by Shafer and Vovk (2019), allows us to test forecasters who behave opportunistically, forecasting new events (hurricanes, sporting events, political outcomes, etc.) as they come along, without comprehensive well-defined models at the outset. Here, however, I will emphasize testing protocols that represent statistical models.

Section 4.1 introduces protocols for testing a single probability distribution. Section 4.2 introduces protocols for testing statistical models and fleshes out the notion of a $(1/\alpha)$-warranty. Section 4.3 discusses how these ideas apply to non-parametric estimation by least squares.

### *4.1. Testing protocols*

We first formalize Section 2's method of testing a probability distribution $P$ for a phenomenon $Y$ that takes values in a set $\mathcal{Y}$. Here, because we have only one probability distribution rather than a statistical model consisting of many candidate probability distributions, we can identify the statistician with the player Sceptic. Sceptic plays against Reality as follows.

**Protocol 1. Testing a probability distribution**

```
    Sceptic announces $S : \mathcal{Y} \to [0, \infty)$ such that $\mathbf{E}_P(S) = 1$.
    Reality announces $y \in \mathcal{Y}$.
    $\mathcal{K} := S(y)$.
```

Like all testing protocols considered in this paper, this is a perfect information protocol; the players move sequentially and each sees the other's move as it is made.

Because $y$ can be multi-dimensional, Protocol 1 can be used to test a probability distribution $P$ for a stochastic process $Y = (Y_1, \dots, Y_N)$. See Shafer and Vovk (2019) for expositions that emphasize processes that continue indefinitely instead of stopping at a non-random time $N$. Often, however, the probability distribution for a stochastic process represents the hypothesis that no additional information we obtain as the process unfolds can provide further help predicting it — more precisely, that no information available at the point when we have observed $y_1, \dots, y_{n-1}$ can enable us to improve on $P$'s conditional probabilities given $y_1, \dots, y_{n-1}$ for predicting $y_n, \dots, y_N$. To test this hypothesis, we may use a perfect-information protocol in which Sceptic observes the $y_n$ step by step:

**Protocol 2. Testing a stochastic process**

```
$\mathcal{K}_0 := 1$.
FOR $n = 1, 2, \dots, N$:
    Sceptic announces $S_n : \mathcal{Y} \to [0, \infty)$ such that
            $\mathbf{E}_P(S_n(Y_n)|y_1, \dots, y_{n-1}) = \mathcal{K}_{n-1}$.
    Reality announces $y_n \in \mathcal{Y}$.
    $\mathcal{K}_n := S_n(y_n)$.
```

The condition of perfect information requires only that each player sees the others' moves as they are made. Some or all of the players may receive additional information as play proceeds.

Sceptic can make any bet against $P$ in Protocol 2 that he can make in Protocol 1. Indeed, for any payoff $S : \mathcal{Y}^N \to [0, \infty)$ such that $\mathbf{E}_P(S) = 1$, Sceptic can play so that $\mathcal{K}_N = S(y_1, \dots, y_N)$; on the $n$th round, he makes the bet $S_n$ given by $S_n(y) := \mathbf{E}_P(S(y_1, \dots, y_{n-1}, y, Y_{n+1}, \dots, Y_N)|y_1, \dots, y_{n-1}, y)$. He can also make bets taking additional information into account.

Many or most statistical models also use additional information (a.k.a., independent variables) to make the probability predictions about a sequence $y_1, \dots, y_N$. We can bring this option into the sequential betting picture by having Reality announce a signal $x_n$ at the beginning of each round and by supplying the protocol with a probability distribution $P_{x_1,y_1\ldots,x_{n-1},y_{n-1},x_n}$ for each round $n$ and each possible sequence of signals and outcomes $x_1, y_1 \ldots, x_{n-1}, y_{n-1}, x_n$ that might precede Sceptic's move on that round:

**Protocol 3. Testing a model with an independent variable**

```
$\mathcal{K}_0 := 1$.
FOR $n = 1, 2, \ldots, N$:
    Reality announces $x_n \in \mathcal{X}$.
    Sceptic announces $S_n : \mathcal{Y} \to [0, \infty)$ such that
            $\mathbf{E}_{P_{x_1,y_1\ldots,x_{n-1},y_{n-1},x_n}}(S_n) = \mathcal{K}_{n-1}$.
    Reality announces $y_n \in \mathcal{Y}$.
    $\mathcal{K}_n := S_n(y_n)$.
```

A simpler protocol is obtained when we drop the assumption that probability distributions are specified at the outset and introduce instead a player, say Forecaster, who decides on them as play proceeds:

**Protocol 4. Testing a forecaster**

```
$\mathcal{K}_0 := 1$.
FOR $n = 1, 2, \ldots, N$:
    Reality announces $x_n \in \mathcal{X}$.
    Forecaster announces a probability distribution $P$ on $\mathcal{Y}$.
    Sceptic announces $S_n : \mathcal{Y} \to [0, \infty)$ such that $\mathbf{E}_P(S_n) = \mathcal{K}_{n-1}$.
    Reality announces $y_n \in \mathcal{Y}$.
    $\mathcal{K}_n := S_n(y_n)$.
```

This protocol allows us to test forecasters who give probabilities for sequences of events without using probability distributions or statistical models fixed at the outset. This includes both weather forecasters who use physical models and forecasters of sporting and electoral outcomes who invent and tinker with models as they go along. Although there is no comprehensive probability distribution or statistical model to test in these cases, we can still rely on the intuition that the forecaster is discredited if Sceptic manages to multiply the capital he risks by a large factor. This intuition is supported by the theory developed by Shafer and Vovk (2019), where it is shown that Sceptic can multiply the capital he risks by a large factor if the probability forecasts actually made do not agree with outcomes in ways that standard probability theory predicts. The reliance on forecasts actually made, without any attention to other aspects of any purported comprehensive probability distribution, makes this approach *prequential* in the sense developed by Dawid (1984).

### 4.2. Statistical models as testing protocols

Now let us turn to testing protocols that represent parametric statistical models. Here the statistician is distinct from Sceptic. Sceptic is a player in the perfect-information game, but the statistician sees only the outcomes.

Setting stochastic processes and signals aside for the sake of simplicity, consider this protocol for $N$ independent and successive observations from a parametric model $(P_\theta)_{\theta\in\Theta}$.

```
**Protocol 5. Independent observations from** $(P_\theta)_{\theta\in\Theta}$
$\mathcal{K}_0 := 1$.
Reality announces $\theta \in \Theta$.
FOR $n = 1, 2, \ldots, N$:
    Sceptic announces $S_n : \mathcal{Y} \to [0, \infty)$ such that $\mathbf{E}_{P_\theta}(S_n) = \mathcal{K}_{n-1}$.
    Reality announces $y_n \in \mathcal{Y}$.
    $\mathcal{K}_n := S_n(y_n)$.
```

The statistician sees neither Reality's move $\theta$ nor Sceptic's moves $S_1, \ldots, S_N$. She sees only the outcomes $y_1, \ldots, y_N$.

Because $\theta$ is announced to Sceptic at the outset, a strategy $\mathcal{S}$ for Sceptic that uses only the information provided by Reality's moves can be thought of as a collection of strategies, one for each $\theta \in \Theta$. The strategy for $\theta$, say $\mathcal{S}^\theta$, specifies Sceptic's move $S_n$ as a function of $y_1, \ldots, y_{n-1}$. This makes Sceptic's final capital a function of $\theta$ and the observations $y_1, \ldots, y_N$. Let us write $\mathcal{K}_{\mathcal{S}}(\theta)$ for this final capital, leaving the dependence on $y_1, \ldots, y_N$ implicit.

Sceptic is a creation of the statistician's imagination and therefore subject to the statistician's direction. Suppose the statistician directs Sceptic to play a particular strategy that uses only Reality's moves. Then, after observing $y_1, \ldots, y_N$, the statistician can calculate Sceptic's final capital is a function of $\theta$, say $\mathcal{K}(\theta)$. Let us call $\mathcal{K}(\theta)$ the statistician's *betting score against the hypothesis* $\theta$. We interpret it just as we interpreted betting scores in Section 2. The statistician doubts that Sceptic has multiplied his initial unit capital by a large factor, and so he thinks that the hypothesis $\theta$ has been discredited if $\mathcal{K}(\theta)$ is large. This way of thinking also leads us to betting scores against composite hypotheses and to a notion of *warranty* analogous to the established notion of confidence.

(a) For each composite hypothesis $\Theta_0 \subseteq \Theta$, $\mathcal{K}(\Theta_0) := \inf\{\mathcal{K}(\theta) | \theta \in \Theta_0\}$ is a *betting score against* $\Theta_0$. When $\mathcal{K}(\Theta_0)$ is large, all the elements of $\Theta_0$ are discredited and hence $\Theta_0$ itself is discredited.

(b) For each $\alpha > 0$,

$$W_{1/\alpha} := \left\{\theta \in \Theta \mid \mathcal{K}(\theta) < \frac{1}{\alpha}\right\}$$

is a $(1/\alpha)$-*warranty set*. This is the set of possible values of $\theta$ that have not been discredited at level $1/\alpha$. We say that the statistician's choice of $S$ has given her a $(1/\alpha)$ *warranty* for this set.

The notion of a warranty was already developed in (Vovk, 1993, Section 7). The intuition can be traced back at least to Schnorr (1971) and Levin (1976).

For small $\alpha$, the statistician will tend to believe that the true $\theta$ is in $W_{1/\alpha}$. For example, she will not expect Sceptic to have multiplied his capital by 1000 and hence will believe that $\theta$ is in $W_{1000}$. But this belief is not irrefutable. If she obtains strong enough evidence that $\theta$ is not in $W_{1000}$, she may conclude that Sceptic actually did multiply his capital by 1000 using $\mathcal{S}$. See Fraser et al. (2018) for examples of outcomes that cast doubt on confidence statements and would also cast doubt on warranties.

Every $(1-\alpha)$-confidence set has a $(1/\alpha)$-warranty. This is because a $(1-\alpha)$-confidence set is specified by testing each $\theta$ at level $\alpha$; the $(1-\alpha)$-confidence set consists of the $\theta$ not rejected. When $\mathcal{S}$ makes the all-or-nothing bet against $\theta$ corresponding to the test used to form the confidence set, $K(\theta) < 1/\alpha$ if and only if $\theta$ was not rejected, and hence $W_{1/\alpha}$ is equal to the confidence set.

Warranty sets are nested: $W_{1/\alpha} \subseteq W_{1/\alpha'}$ when $\alpha \leq \alpha'$. Standard statistical theory also allows nesting; sets with different levels of confidence can be nested. But the different confidence sets will be based on different tests (Cox, 1958; Xie and Singh, 2013). The $(1/\alpha)$-warranty sets for different $\alpha$ all come from the same strategy for Sceptic.

Instead of stopping the protocol after some fixed number of rounds, the statistician may stop it when she pleases and adopt the $(1/\alpha)$-warranty sets obtained at that point. As we learned in Section 2, the intuition underlying betting scores supports such optional continuation; multiplying the money you risk by a large factor discredits a probabilistic hypothesis or a probability forecaster no matter how you decide to bet and no matter how long you persist in betting. The only caveat is that we cannot pretend to have stopped before we actually did (Shafer and Vovk, 2019, Ch. 11). This contrasts with confidence intervals; if we continually calculate test results and the corresponding confidence intervals as we make more and more observations, the multiple testing vitiates the confidence coefficients and so may be called "sampling to a foregone conclusion" (Cornfield, 1966; Shafer et al., 2011). The most important exceptions are the "confidence sequences" that can be obtained from the sequential probability ratio test (Lai, 2009). Because they are derived from products of successive likelihood ratios that can be interpreted as betting scores, these confidence sequences can be understood as sequences of warranty sets.

How should the statistician choose the strategy for Sceptic? An obvious goal is to obtain small warranty sets. But a strategy that produces the smallest warranty set for one $N$ and one warranty level $1/\alpha$ will not generally do so for other values of these parameters. So any choice will be a balancing act. How to perform this balancing act is an important topic for further research (Grünwald et al., 2019).

### 4.3. Non-parametric least squares

Consider the statistical hypothesis that observations $e_1, e_2, \ldots$ are drawn independently from an unknown probability distribution $P$ on $[-1, 1]$ that has mean zero. How can we test this hypothesis by betting?

If $P$ were fully specified, then we could use a version of Protocol 2; on the $n$th round Sceptic would be allowed to buy any nonnegative payoff $S_n(e_n)$ such that $\mathbf{E}_P(S_n(e_n)) = \mathcal{K}_{n-1}$. But the hypothesis we want to test specifies expected values only for linear functions of $e_n$: $\mathbf{E}_P(ae_n + b) = b$. So we can only authorize Sceptic to buy payoffs of the form $ae_n + \mathcal{K}_{n-1}$ that are nonnegative whenever $e_n \in [-1, 1]$. This leads us to the following testing protocol.

```
Protocol 6. Betting on successive bounded outcomes
$\mathcal{K}_0 := 1$.
FOR $n = 1, 2, \ldots$:
    Sceptic announces $a_n \in [-\mathcal{K}_{n-1}, \mathcal{K}_{n-1}]$.
    Reality announces $e_n \in [-1, 1]$.
    $\mathcal{K}_n := \mathcal{K}_{n-1} + a_n e_n$.
```

In (Shafer and Vovk, 2019, Section 3.3), it is shown that Sceptic has a strategy in Protocol 6, based on Hoeffding's inequality, that guarantees $\mathcal{K}_n \geq 20$ for every $n$ such that $|\overline{e}_n| > 2.72/\sqrt{n}$, where $\overline{e}_n$ is the average of $e_1, \ldots, e_n$. If Sceptic plays this strategy and eventually reaches an $n$ for which $|\overline{e}_n| > 2.72/\sqrt{n}$, then he can claim a betting score of 20 against the hypothesis. Had we specified a probability distribution $P$ on $[-1, 1]$ with mean zero and allowed Sceptic to choose on each round any nonnegative payoff with expected value under $P$ equal to his current capital, then he could have chosen a particular value $N$ large enough for the central limit theorem to be effective and claimed a betting score of 20 if $|\overline{e}_N| > 2.0/\sqrt{N}$. But this would work only for the particular value $N$.

Now suppose that the $e_n$ are errors for successive measurements of a quantity $\mu$. As in Section 4.2, assume that the statistician stands outside the game and sees only $y_1, y_2 \ldots$. She does not see $\mu$ or the errors $e_1, e_2, \ldots$. Then we have this protocol.

```
Protocol 7. Estimating $\mu$
$\mathcal{K}_0 := 1$.
Reality announces $\mu \in \mathbb{R}$.
FOR $n = 1, 2, \ldots$:
    Sceptic announces $a_n \in [-\mathcal{K}_{n-1}, \mathcal{K}_{n-1}]$.
    Reality announces $e_n \in [-1, 1]$ and sets $y_n := \mu + e_n$.
    $\mathcal{K}_n := \mathcal{K}_{n-1} + a_n e_n$.
```

Now the strategy for Sceptic that guarantees $\mathcal{K}_n \geq 20$ for every $n$ such that $|\overline{e}_n| > 2.72\sqrt{n}$ can be used to obtain warranties for $\mu$. After 100 measurements, for example, it gives a 20-warranty that $\mu$ is in $\overline{y}_{100} \pm 0.272$, where $\overline{y}_{100}$ is the average of $y_1, \ldots, y_{100}$.

The statistician will know the betting score that Sceptic has achieved only as a function of $\mu$. But a meta-analyst, imagining that Sceptic has used his winnings from each study in the next study, can multiply the functions of $\mu$ obtained from multiple stuides to obtain warranties about $\mu$ that may be more informative and authoritative than those from the individual studies.

Averaging measurements of a single quantity to estimate the quantity measured is the most elementary instance of estimation by least squares. The ideas developed here extend to the general theory of estimation by least squares, in which $\mu$ is multi-dimensional and multi-dimensional signals $x_1, x_2, \ldots$ are used. An asymptotic theory with this generality, inspired by Lai and Wei (1982), is developed in (Shafer and Vovk, 2019, Section 10.4).

## 5. Conclusion

The probability calculus began as a theory about betting, and its logic remains the logic of betting, even when it serves to describe phenomena. But in their quest for the appearance of objectivity, mathematicians have created a language (likelihood, significance, power, p-value, confidence) that pushes betting into the background.

This deceptively objective statistical language can encourage overconfidence in the results of statistical testing and neglect of relevant information about how the results are obtained. In recent decades this problem has become increasingly salient, especially in medicine and the social sciences, as numerous influential statistical studies in these fields have turned out to be misleading.

In 2016, the American Statistical Association issued a statement listing common misunderstandings of p-values and urging full reporting of searches that produce p-values (Wasserstein and Lazar, 2016). Many statisticians fear, however, that the situation will not improve. Most dispiriting are studies showing that both teachers of statistics and scientists who use statistics are apt to answer questions about the meaning of p-values incorrectly (McShane and Gal, 2017; Gigerenzer, 2018). Andrew Gelman and John Carlin argue persuasively that the most frequently proposed solutions (better exposition, confidence intervals instead of tests, practical instead of statistical significance, Bayesian interpretation of one-sided p-values, and Bayes factors) will not work (Gelman and Carlin, 2017). The only solution, they contend, is “to move toward a greater acceptance of uncertainty and embracing of variation” (p. 901).

In this context, the language of betting emerges as an important tool of communication. When statistical tests and conclusions are framed as bets, everyone understands their limitations. Great success in betting against probabilities may be the best evidence we can have that the probabilities are wrong, but everyone understands that such success may be mere luck. Moreover, candor about the betting aspects of scientific exploration can communicate truths about the games scientists must and do play — honest games that are essential to the advancement of knowledge.

This paper has developed new ways of expressing statistical results with betting language. The basic concepts are *bet* (not necessarily all-or-nothing), *betting score* (equivalent to likelihood ratio when the bets offered define a probability distribution), *implied target* (an alternative to power), and $(1/\alpha)$-*warranty* (a generalization of $(1-\alpha)$-confidence). Substantial research is needed to apply these concepts to complex models, but their greatest utility may be in communicating the uncertainty of simple tests and estimates.

## 6. Appendix: Situating testing by betting

*Disclaimers.* My theme has been that we can communicate statistical conclusions and their uncertainty more effectively with betting scores than with p-values. This is not to say that well trained statisticians are unable to use p-values effectively. We have been using them effectively for two centuries.

I have emphasized that testing by betting is not a chapter in decision theory, because tests can use amounts of money so small that no one cares about them. The betting is merely to make a point, and play money would do. This is not to say that decision theory is an unimportant chapter in statistical methodology. Many statistical problems do involve decisions for which the utilities and probabilities required by various decision theories are available. These theories include the Neyman-Pearson theory and Bayesian theory. These theories do not use p-values, and so replacing p-values by betting scores would not affect them.

*Are the probabilities tested subjective or objective?* The probabilities may represent someone's opinion, but the hypothesis that they say something true about the world is inherent in the project of testing them.

*Are the probabilities being tested frequencies?* Sometimes. In Protocol 5, Sceptic can select any particular event to which $P$ assigns a probability and adopt a strategy that will produce a large betting score unless Reality makes the frequency of the event approximate that probability. But only the most salient probabilities will be tested, and as Abraham Wald pointed out in the 1930s, only a countable number of them could be tested (Bienvenu et al., 2009). So the identification of $P$'s probabilities with frequencies is always approximate and usually hypothetical. The connection with frequencies is even more tenuous when the theory tested involves limited betting offers, as in non-parametric and other imprecise-probability models.

*Does testing by betting extend from probabilities to imprecise probabilities?* Yes. As explained by Augustin et al. (2014), imprecise probabilities can be expressed in terms of upper and lower prices for payoffs. If the payoffs depend on events that will be decided, so that bets can be settled, then we can test the concordance of the prices with actual outcomes by interpreting the prices as betting offers. The only difference from the case of a complete probability distribution is that there are fewer betting offers. If we say that probabilities are imprecise whenever there are fewer betting offers than would be provided by a complete probability distribution, then Protocols 3, 4, 6, and 7 are all examples of testing imprecise probabilities. For an abstract discussion of testing imprecise probabilities by betting, see (Shafer and Vovk, 2019, Ch. 6).

*How is testing by betting related to Bruno de Finetti's theory of probability?* In de Finetti's picture, an individual who has evidence about a hypothesis or a question of fact expresses the strength of that evidence by betting odds and prices for payoffs that depend on the truth of the hypothesis or the answer to the question (de Finetti, 1970). In the case of statistical hypotheses, this becomes, curiously, a theory about bets that are never settled. The odds for the statistical hypothesis change as evidence accumulates, but we usually never decide for certain whether the hypothesis is true, and so there is no settling up.

Testing by betting, in contrast, is concerned with bets that are settled. It uses *how a bet came out* as a measure of *evidence about probabilities.* This is a very different undertaking than de Finetti's, and it would not be productive to try to explain the one undertaking in terms of the other.

Another contrast emerges in protocols in which Forecaster chooses and announces betting offers. De Finetti's viewpoint was that of Forecaster. Most authors on imprecise probabilities, including Peter Walley (Walley, 1991), have followed de Finetti in this respect. Testing by betting emphasizes the viewpoint of Sceptic, who decides how to bet.

Some readers have asked how the expectation that Sceptic not obtain a large betting score is related to de Finetti's condition of coherence. The two are not closely related. Coherence is about Forecaster's behavior: he should not make offers that allow Sceptic to make money no matter what Reality does. This condition is met by all the protocols in this paper. The expectation that Sceptic not obtain a large betting score is about Reality's behavior. When Reality violates this expectation, we doubt whether the betting offers are a valid description of Reality.

*Why is the proposal to test by betting better than other proposals for remedying the misuse of p-values?* Many authors have proposed remedying the misuse of p-values by supplementing them with additional information (Wasserstein et al., 2019; Mayo, 2018). Sometimes the additional information involves Bayesian calculations (Bayarri et al., 2016; Matthews, 2018). Sometimes it involves likelihood ratios (Colquhoun, 2019). Sometimes it involves attained power (Mayo and Spanos, 2006).

I find nearly all these proposals persuasive as ways of correcting the misunderstandings and misinterpretations to which p-values are susceptible. Each of them might be used by a highly trained mathematical statistician to explain what has gone wrong to another highly trained mathematical statistician. But adding more complexity to the already overly complex idea of a p-value may not help those who are not specialists in mathematical statistics. We need strategies for communicating with millions of people. Worldwide, we teach p-values to millions every year, and hundreds of thousands of them may eventually use statistical tests in one way or another.

The strongest argument for betting scores as a replacement for p-values is its simplicity. I do not know any other proposal that is equally simple.

*Is testing by betting a novel idea?* The idea of testing purported probabilities by betting is part of our culture. But it is almost always below the surface in the mathematics of probability and statistics. Some authors allude to betting from time to time, but they usually mention only all-or-nothing bets. Betting never takes center stage, even though it is central to everyone's intuitions about probability. Ever since Jacob Bernoulli, we have downplayed the betting intuition in order to make probabilities look more objective — more scientific.

Betting did come to the surface in the work of Jean Ville (Ville, 1939, pp. 87–89). Richard von Mises's principle of the impossibility of a gambling system (von Mises, 1928, p. 25) said that a strategy for selecting throws on which to bet should not improve a gambler's chances. Ville replaced this with a stronger and more precise condition: the winnings from a strategy that risks only a fixed amount of money should not increase indefinitely, even if the strategy is allowed to vary the amount bet and the side on which to bet. As this wording suggests, Ville was concerned with what betting strategies accomplish in an infinite sequence of play. He did not consider statistical testing, and his work has had little or no influence on mathematical statistics.

Many statisticians, including Fisher, have advocated using likelihood as a direct measure of evidence. In his *Statistical Methods and Scientific Inference* (Fisher, 1956), Fisher suggested that in most cases the plausible values of a parameter indexing a class of probability distributions are those for which the likelihood is at least (1/15)th of its maximum value. On pp. 71–73 of the first edition, he used diagrams to show "outside what limits the likelihood falls to levels at which the corresponding values of the parameter become implausible." In these diagrams,

> ...zones are indicated showing the limits within which the likelihood exceeds 1/2, 1/5, and 1/15 of the maximum. Values of the parameter outside the last limit are obviously open to grave suspicion.

Later authors, including Edwards (1972) and Royall (1997), published book-length arguments for using likelihood ratios to measure the strength of evidence, and articles supporting this viewpoint continue to appear. I have not found in this literature any allusion to the idea that a likelihood ratio measures the success of a bet against a null hypothesis.

Because Ville introduced martingales into probability theory, we might expect that statisticians who use martingales would recognize the possibility of interpreting nonnegative martingales directly as tests. They know that when $P$ and $Q$ are probability distributions for a sequence $Y_1, Y_2, \ldots$, the sequence of likelihood ratios

$$1, \frac{Q(Y_1)}{P(Y_1)}, \frac{Q(Y_1, Y_2)}{P(Y_1, Y_2)}, \ldots \tag{7}$$

is a nonnegative martingale, and they would see no novelty in the following observations:

(a) If I am allowed to bet on $Y_1, \ldots, Y_n$ at rates given by $P$, then I can buy the payoff $Q(Y_1, \ldots, Y_n)/P(Y_1, \ldots, Y_n)$ for one monetary unit.

(b) If I bet sequentially, on each $Y_k$ at rates given by $P(Y_k|y_1, \ldots, y_{k-1})$ when $y_1, \ldots, y_{k-1}$ are known, and I always use my winnings so far to buy that many units of $Q(Y_k|y_1, \ldots, y_{k-1})/P(Y_k|y_1, \ldots, y_{k-1})$, then (7) will be the capital process resulting from my strategy.

But they also know that Joseph L. Doob purified the notion of a martingale of its betting heritage when he made it a technical term in modern probability theory. Martingales are now widely used in sequential analysis, time series, and survival analysis (Aalen et al., 2009; Lai, 2009), but I have not found in the statistical literature any use of the idea that the successive realized values of a martingale already represent, without being translated into the language of p-values and significance levels, the cumulative evidential value of the outcomes of a betting strategy.

## 7. Acknowledgements

The game-theoretic foundation for probability on which this paper relies is based primarily on the insights of Vladimir Vovk, and the paper has been influenced in numerous ways by my collaboration with him over the past several decades.

Conversations with many others over the past several years have also influenced the paper. The paper was inspired by conversations about game-theoretic testing and meta-analysis with Peter Grünwald, Wouter Koolen, and Judith ter Schure at the Centrum Wiskunde & Informatica in Amsterdam in December 2018. Also especially important were conversations with Gert de Cooman and Jasper De Bock and their students at the University of Ghent, with Harry Crane, Jacob Feldman, Robin Gong, Barry Loewer, and others in Rutgers University's seminar on the Foundations of Probability, and with Jason Klusowski, William Strawderman, and Min-ge Xie in the seminar of the Statistics Department at Rutgers.

Many others have provided useful feedback after the paper was first drafted, including John Aldrich, Peter Carr, Steve Goodman, Prakash Gorroochurn, Sander Greenland, Alan Hájek, David Madigan, Deborah Mayo, Rohit Parikh, Arthur Paul Pedersen, Teddy Seidenfeld, Stephen Senn, Nozer Singpurwalla, Mike Smithson, Aris Spanos, Matthias Troffaes, Vladimir Vovk, Conor Mayo-Wilson, and several anonymous referees.

## References

Aalen, O. O., P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding (2009). History of applications of martingales in survival analysis. *Electronic Journal for History of Probability and Statistics 5*(1).

Amrhein, V., S. Greenland, and B. McShane et al. (2019). Retire statistical significance. *Nature 567*, 305–307.

Augustin, T., F. P. A. Coolen, G. de Cooman, and M. C. M. Troffaes (Eds.) (2014). *Introduction to Imprecise Probabilities.* Wiley.

Bayarri, M. J., D. J. Benjamin, J. O. Berger, and T. M. Sellke (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology 72*, 90–103.

Bennett, J. H. (Ed.) (1990). *Statistical inference: Selected correspondence of R. A. Fisher.* Oxford: Clarendon.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association 33*(203), 526–536.

Bienvenu, L., G. Shafer, and A. Shen (2009). On the history of martingales in the study of randomness. *Electronic Journal for History of Probability and Statistics 5*(1).

Breiman, L. (1961). Optimal gambling systems for favorable games. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1 (Contributions to the Theory of Statistics), Berkeley, CA, pp. 65–78. University of California Press.

Cesa-Bianchi, N. and G. Lugosi (2006). *Prediction, Learning, and Games.* Cambridge, UK: Cambridge University Press.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science 349*(6251), 943.

Colquhoun, D. (2019). The false positive risk: A proposal concerning what to do about p-values. *The American Statistician 73*(sup1), 192–201.

Cornfield, J. (1966). A Bayesian test of some classical hypotheses — with applications to sequential clinical trials. *Journal of the American Statistical Association 61*, 577–594.

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory.* New York: Wiley. Second edition in 2006.

Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics 29*, 357–372.

Cready, W. M. (2019, August). Complacency at the gates: A field report on the non-impact of the ASA Statement on Statistical Significance and P-Values on the broader research community. *Significance 16*(4), 18–19.

Cready, W. M., J. He, W. Lin, C. Shao, D. Wang, and Y. Zhang (2019). Is there a confidence interval for that? A critical examination of null outcome reporting in accounting research. Available at SSRN: https://ssrn.com/abstract=3131251 or http://dx.doi.org/10.2139/ssrn.3131251.

Dawid, A. P. (1984). Statistical theory: The prequential approach (with discussion). *Journal of the Royal Statistical Society. Series A 147*(2), 278–292.

de Finetti, B. (1970). *Teoria Delle Probabilità.* Turin: Einaudi. An English translation, by Antonio Machi and Adrian Smith, was published as *Theory of Probability* by Wiley (London, England) in two volumes in 1974 and 1975.

Dempster, A. P. (1997). The direct use of likelihood for significance testing. *Statistics and Computing 7*(4), 247–252. This article is followed on pages 253–272 by a related article by Murray Aitkin and further discussion by Dempster, Aitkin, and Mervyn Stone. It originally appeared on pages 335–354 of **?**) along with discussion by George Barnard and David Cox.

Edwards, A. W. F. (1972). *Likelihood. An account of the statistical concept of likelihood and its application to scientific inference.* Cambridge: Cambridge University Press.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London (A) 222*, 309–368.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd. The thirteenth edition appeared in 1958.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd. Subsequent editions appeared in 1959 and 1973.

Fraser, D. A. S., N. Reid, and W. Lin (2018). When should modes of inference disagree? Some simple but challenging examples. *Annals of Applied Statistics 12*(2), 750–770.

Gelman, A. and J. Carlin (2017). Some natural solutions to the p-value communication problem—and why they won't work. *Journal of the American Statistical Association 112*(519), 899–901.

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science 1*(2), 198–218.

Grünwald, P. D., R. d. Heide, and W. M. Koolen (2019). Safe testing. arXiv:1906.07801 [math.ST].

Harvey, C. R. (2017). The scientific outlook in financial economics. *Journal of Finance 72*(4), 1399–1440.

Kelly Jr., J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal 35*(4), 917–926.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

Lai, T. L. (2009). History of martingales in sequential analysis and time series. *Electronic Journal for History of Probability and Statistics 5*(1).

Lai, T. L. and C. Z. Wei (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics 10*(1), 154–166.

Levin, L. A. (1976). Uniform tests of randomness (in Russian). *Doklady Akademii Nauk SSSR 227*(1), 33–35. `http://mi.mathnet.ru/dan40194`.

Luenberger, D. G. (2014). *Investment Science* (second ed.). New York: Oxford University Press.

Madigan et al., D. (2014). A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Applications 1*, 11–39.

Matthews, R. A. (2018). Beyond 'significance': principles and practice of the analysis of credibility. *Royal Society Open Science 5*, 171047. http://dx.doi.org/10.1098/rsos.171047.

Mayo, D. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars.* Cambridge.

Mayo, D. G. and A. Spanos (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the Philosophy of Science 57*, 323–357.

McShane, B. B. and D. Gal (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association 112*(519), 885–895.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences 236*(767), 333–380.

Neyman, J. and E. S. Pearson (1928). On the use and interpretation of certain test criteria. *Biometrika 20A*, 175–240, 263–295.

Royall, R. M. (1997). *Statistical Evidence: A Likelihood Paradigm.* London: Chapman & Hall.

Schnorr, C.-P. (1971). *Zufälligkeit und Wahrscheinlichkeit. Eine algorithmische Begründung der Wahrscheinlichkeitstheorie.* Springer.

Schuemie et al., M. J. (2018). Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philosophical Transactions of the Royal Society, Series A 376.*

Senn, S. (2011). You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets, and Morals 2*(42), 48–66.

Shafer, G. (2019). On the nineteenth century origins of significance testing and p-hacking. Working Paper 55, `www.probabilityandfinance.com`.

Shafer, G., A. Shen, N. Vershchagin, and V. Vovk (2011). Test martingales, Bayes factors, and p-values. *Statistical Science 26*, 84–101.

Shafer, G. and V. Vovk (2019). *Game-Theoretic Foundations for Probability and Finance.* Hoboken, New Jersey: Wiley.

ter Schure, J. and P. Grünwald (2019). Accumulation bias in meta-analysis: The need to consider time in error control. arXiv:1905.13494 [stat.ME].

Ville, J. (1939). *Étude critique de la notion de collectif.* Paris: Gauthier-Villars.

von Mises, R. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit.* Wien: Springer.

Vovk, V. (1993). A logic of probability, with applications to the foundations of statistics (with discussion). *Journal of the Royal Statistical Society. Series B 55*(2), 317–351.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall.

Wasserstein, R. L. and N. A. Lazar (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician 70*(2), 129–133.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar (2019). Moving to a world beyond "$p < 0.05$". *The American Statistician 73*(sup1), 1–19.

Xie, M.-g. and K. Singh (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review (with discussion). *International Statistical Review 81*(1), 3–77.