Received: 15 March 2019 | Accepted: 11 May 2020

DOI: 10.1111/rssa.12647

Journal of the Royal Statistical Society Series A Statistics in Society

ORIGINAL ARTICLE

Discussion Paper

Testing by betting: A strategy for statistical and scientific communication

Glenn Shafer

Rutgers University, Newark, NJ, USA

Correspondence
Glenn Shafer, Rutgers University, Newark, NJ, USA.
Email: gshafer@business.rutgers.edu

Abstract

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 (Gigerenzer, 2018; McShane & Gal, 2017). 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.

[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]

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:

1. 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.
2. 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.
3. 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.
4. 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 modelling 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:

1. Like individual betting scores, it gives colour to residual uncertainty by evoking our knowledge of gambling and its dangers.
2. Observations together with a strategy for the bettor produce more than one warranty set. They produce a $(1/\alpha)$-warranty set for every α, and these warranty sets are nested.

3. 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.

2.1.1 | 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 a 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_α such that

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

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

2.1.2 | 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 $Y_1, \ldots, Y_N$ 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. If the scientist uses bets that are not all-or-nothing, 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 scepticism 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 (Open Science 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 & 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 Q, the expected gain from $S := Q/P$ is nonnegative: $\mathbf{E}_Q(S) \geq 1$. In fact, as a referee has pointed out, $\mathbf{E}_Q(S) = \mathbf{E}_P(S^2)$ and hence

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

2.2.1 | 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 $\mathbf{E}_Q(\ln(Q/P))$ 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 favour 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 α?

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 & Thomas, 1991), finance theory (Luenberger, 2014) and machine learning (Cesa-Bianchi & 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 & 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 for which Q/P is greatest. Is a bet that risks everything on this single possible outcome a sensible test? If $Q(y)/P(y)$ is huge, much greater than we would need to refute Q, and yet $Q(y)$ 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 favour 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.

2.2.2 | The Neyman–Pearson lemma

In 1928, Jerzy Neyman and E. S. Pearson suggested that for a given significance level α, 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 & 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-α 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

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

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-α 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-α 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 and the fact that $\mathbf{E}_P(Q/P) = 1$, we find that

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

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.

2.2.3 | 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 five 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 favour 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 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 $\mathbf{E}_Q(\ln S)$—that is, a betting score in the ballpark of

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

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:

1. 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 & Carlin, 2017, p. 900).
2. 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).
3. 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 $y = 30$.

1. Statistician A simply calculates a p-value: $P(Y \geq 30) \approx 0.00135$. She concludes that P is strongly discredited.
2. 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.
3. Statistician C uses the bet S given by

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
$\exp\left(\mathbf{E}_Q(\ln S)\right)$	Implied target	S^*
Results		
Actual value of Y	Outcome	y
Factor by which money risked has been multiplied	Betting score	$S(y)$

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

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) ≈ 1.005. She does a little better than this very low target; she multiplies the money she risked by exp (59/200) ≈ 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 favour 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 $y = 16.5$.

1. Statistician A again calculates a p-value: $P(Y \geq 16.5) \approx 0.0495$. She concludes that P is discredited.
2. 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.
3. 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 favours 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 $y = 5$.

1. 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.
2. 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.
3. 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 favours 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 $S(y)$ favours 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 meagreness 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 value 1 or less.) No one has made a convincing case for any particular choice from this infinitude; the choice is fundamentally arbitrary (Shafer & 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 favourite, 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 these data, Fisher tested the bias of the dice in two different ways.

1. First, he performed a χ^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 χ^2 statistic came out 40.748, and he noted that 'the actual chance in this case of χ^2 exceeding 40·75 if the dice had been true is ·00003'.
2. Then he noted that in the 12×26,306 = 315,672 throws of a die there were altogether 106,602 5 and 6 s. 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 2000 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

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

TABLE 2 Making a *p*-value into a betting score

***p*-value**	$\frac{1}{p-\text{value}}$	$\frac{1}{\sqrt{p-\text{value}}}-1$
0.10	10	2.2
0.05	20	3.5
0.01	100	9.0
0.005	200	13.1
0.001	1000	30.6
0.000001	1000,000	999

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 ≈ 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 Φ 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 χ^2 test. Whereas the distribution of the χ^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 α. 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 θ is one of the moves she does not see. The player who bets, whom we call Sceptic, does see θ. A strategy for Sceptic tells him how to move depending on the value of θ. 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 Sceptic be offered bets priced by a given probability distribution P or P_θ, 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 $Y = (Y_1, \ldots, Y_N)$. 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 $y_1, \ldots, y_{n-1}$ can enable us to improve on P's conditional probabilities given $y_1, \ldots, y_{n-1}$ for predicting $y_n, \ldots, y_N$. 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 $n = 1,2,\ldots,N$:

Sceptic announces $S_n: \mathcal{Y} \to [0, \infty)$ such that $\mathbf{E}_P(S_n(Y_n) \mid y_1, \ldots, y_{n-1}) = \mathcal{K}_{n-1}$.

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 $\mathcal{K}_N = S(y_1, \ldots, y_N)$; on the nth round, he makes the bet S_n given by $S_n(y) := \mathbf{E}_P(S(y_1, \ldots, y_{n-1}, y, Y_{n+1}, \ldots, Y_N) \,|\, y_1, \ldots, y_{n-1}, y)$. 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 $y_1, \ldots, y_N$. 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_n 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 θ nor Sceptic's moves $S_1, \ldots, S_N$. She sees only the outcomes $y_1, \ldots, y_N$.

Because θ 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 θ, 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 θ 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 θ, say $\mathcal{K}(\theta)$. Let us call $\mathcal{K}(\theta)$ the statistician's *betting score against the hypothesis* θ. 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 θ 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.

1. For each composite hypothesis $\Theta_0 \subseteq \Theta$, $\mathcal{K}(\Theta_0) := \inf\{\mathcal{K}(\theta) \mid \theta \in \Theta_0\}$ is a *betting score against* Θ_0. When $\mathcal{K}(\Theta_0)$ is large, all the elements of Θ_0 are discredited and hence Θ_0 itself is discredited.
2. 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 θ that have not been discredited at level $1/\alpha$. We say that the statistician's choice of $\mathcal{S}$ 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 α, the statistician will tend to believe that the true θ is in $W_{1/\alpha}$. For example, she will not expect Sceptic to have multiplied his capital by 1000 and hence will believe that θ is in W_{1000}. But this

belief is not irrefutable. If she obtains strong enough evidence that θ is not in W_{1000}, she may conclude that Sceptic actually did multiply his capital by 1000 using 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 θ at level α; the $(1-\alpha)$-confidence set consists of the θ not rejected. When S makes the all-or-nothing bet against θ corresponding to the test used to form the confidence set, $\mathcal{K}(\theta) < 1/\alpha$ if and only if θ 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 & Singh, 2013). The $(1/\alpha)$-warranty sets for different α 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 & 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 nth 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 & 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 μ. 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 μ or the errors $e_1, e_2, \ldots$. Then we have this protocol.

Protocol 7. Estimating μ

$\mathcal{K}_0 := 1$.
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 μ. After 100 measurements, for example, it gives a 20-warranty that μ 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 μ. But a meta-analyst, imagining that Sceptic has used his winnings from each study in the next study, can multiply the functions of μ obtained from multiple studies to obtain warranties about μ 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 μ 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 & 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 & 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 (Gigerenzer, 2018; McShane & Gal, 2017). 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 & 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, candour 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.

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

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

Augustin, T., Coolen, F.P.A., de Cooman, G. & Troffaes, M.C.M. (Eds.) (2014) *Introduction to imprecise probabilities*. Hoboken: Wiley.

Barndorff-Nielson, O., Blaesild, P. & Schou, G. (1974) *Proceedings of Conference on Foundational Questions in Statistical Inference, Aarhus, May 7-12, 1973*. Institute of Mathematics, University of Aarhus.

Bayarri, M.J., Benjamin, D.J., Berger, J.O. & Sellke, T.M. (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., Shafer, G. & Shen, A. (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: University of California Press, pp. 65–78.

Cesa-Bianchi, N. & Lugosi, G. (2006) *Prediction, learning, and games*. Cambridge, UK: Cambridge University Press.

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. & Thomas, J.A. (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) 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., He, J., Lin, W., Shao, C., Wang, D. & Zhang, Y. (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.

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 (Barndorff-Nielsen et al., 1974) 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.

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.

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

Gelman, A. & Carlin, J. (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., Heide, R.D. & Koolen, W.M. (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. & Wei, C.Z. (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* (2nd ed.). New York: Oxford University Press.

Madigan, D., Stang, P.E., Berlin, J.A., Schuemie, M., Overhage, J.M., Suchard, M.A., Dumouchel, B., Hartzema, A.G. & Ryan, P.B. (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: Cambridge University Press.

Mayo, D.G. & Spanos, A. (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. & Gal, D. (2017) Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, 112 (519), 885–895.

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

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

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

Royall, R.M. (1997) *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.

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

Schuemie, M.J., Ryan, P.B., Hripcsa, G., Madigan, D. & Suchard, M.A. (2018) Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philosophical Transactions of the Royal Society, Series A*, 376, 20170356.

ter Schure, J. & Grünwald, P. (2019) Accumulation bias in meta-analysis: The need to consider time in error control. arXiv:1905.13494 [stat.ME].https://doi.org/10.12688/f1000research.19375.1

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. & Vovk, V. (2019) *Game-theoretic foundations for probability and finance*. Hoboken, NJ: Wiley.

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

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

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*. London: Chapman and Hall.

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

Wasserstein, R.L., Schirm, A.L. & Lazar, N.A. (2019) Moving to a world beyond "$p < 0.05$". *The American Statistician*, 73(sup1), 1–9.

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

How to cite this article: Shafer G. Testing by betting: A strategy for statistical and scientific communication. *J R Stat Soc Series A*. 2021;184:407–431. https://doi.org/10.1111/rssa.12647

APPENDIX A

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 & 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 behaviour: 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 behaviour. 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 (Mayo, 2018; Wasserstein et al., 2019). 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 & 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 centre 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:

1. 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.
2. If I bet sequentially, on each Y_k at rates given by $P(Y_k \mid y_1, \ldots, y_{k-1})$ when $y_1, \ldots, y_{k-1}$ are known, and I always use my winnings so far to buy that many units of $Q(Y_k \mid y_1, \ldots, y_{k-1})/P(Y_k \mid y_1, \ldots, y_{k-1})$, 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.

DOI: 10.1111/rssa.12648

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

DISCUSSION CONTRIBUTIONS

Proposer of the vote of thanks to Glenn Shafer and contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication'

Philip Dawid
University of Cambridge, Cambridge, UK

Correspondence: Philip Dawid, University of Cambridge, Cambridge, UK.
Email: apd@statslab.cam.ac.uk

I very much like Glenn Shafer's introduction of a betting paradigm for statistical communication, and hope it will catch on. However, I foresee many obstacles to its wide adoption. It takes as given that betting language and behaviour are intuitive and natural. While this may be so for the racehorse fancier or poker player (see Duke (2018) for a nice account of how to repurpose such skills for everyday decision making), I fear that the statistical and scientific communities, already 're-educated' to believe six impossible things before breakfast, may have more trouble with it. There is also the problem of the dubious moral connotations of betting.

The principal novelty of the betting approach is to replace the all-or-nothing bet underlying a traditional hypothesis test with a more general payoff function—together with the insistence (a version of the likelihood principle) that it is the actual value of the eventual payoff, without any further modification to account for its provenance, that carries the evidential meaning of the data. Shafer has made convincing arguments for the benefits that this extension could bring, to both the theory and the communication of Statistics. I particularly like his extension of the Neyman–Pearson lemma, with its simple proof.

However, while the classical approach for testing P has to make a choice among many possible rejection regions, the betting approach faces the still more difficult choice among the great multitude of available bets. When a specific alternative Q is taken seriously, the likelihood ratio supplies an optimal answer in both cases. But when there is no such Q, we have an embarrassment of riches. One possible ploy would be to place a family of bets, and expect to control the maximum score attained—after all, if P is a valid description, none of these is expected to be large. But could we then treat the maximum score, unadorned, as a measure of the evidence against P, or would it need some sort of multiplicity adjustment, *à la* Bonferroni?—perhaps now needing to take account of the joint distribution of the scores.

While the betting story is fairly intuitive when we are testing a single probability distribution, when we move on to inference for a parametric family it may be harder to motivate. In particular, what has happened, in these extensions, to the useful concepts of implied alternative hypothesis and implied target value? And does a $(1/\alpha)$-warranty set have any property analogous to the coverage property of a classical confidence interval?

In the parametric case, the problem of choosing between bets becomes still more difficult, since Statistician now needs to specify a bet for each value of the parameter. As currently formulated, there do not seem to be any constraints on these choices, but it would seem reasonable to expect to tie this collection together, somehow. Indeed, such a connexion is implicit in the formulation of Shafer's Protocol 7, which can be equivalently described as follows:

Protocol 7′

$\mathcal{K}_0 := 1.$

Reality announces $\mu \in \mathbb{R}$.

FOR n=1,2,…:

Sceptic announces $a_n \in [-\mathcal{K}_{n-1}, \mathcal{K}_{n-1}]$.

Reality announces $y_n \in [\mu - 1, \mu + 1]$.

$\mathcal{K}_n := \mathcal{K}_{n-1} + a_n(y_n - \mu)$.

Let a strategy for Protocol 6 (e.g. that described just below it in the paper) be expressed as

$$a_n = \phi_n(e_1, \ldots, e_{n-1}).$$

Then for Protocol 7′ let Statistician instruct Sceptic to play

$$a_n = \phi_n(y_1 - \mu, \ldots, y_{n-1} - \mu). \tag{1}$$

(This will necessarily satisfy the constraint $a_n \in [-\mathcal{K}_{n-1}, \mathcal{K}_{n-1}]$, even though Statistician is not aware of $\mathcal{K}_{n-1}$). The strategy envisaged for Protocol 7 is of this form.

We see that, by involving a common function ϕ_n, Equation (1) makes a simple link between the immediate payoff functions, $a_n(y_n - \mu)$, for different values of μ. This construction could readily be extended to group-structured models, such as the 'structural models' of Fraser (1968), once again tying together the payoff functions across parameter values. But are there any principles that might govern this for other kinds of models? In particular, what about exponential families? Are there analogues of concepts such as uniformly most powerful tests?

One really nice aspect of the betting formulation is how it can handle very sparsely specified sequential problems, and in particular how it allows 'stopping when ahead' without requiring any adjustment—a property it shares with likelihood inference. That does however assume that all the data generated are presented. A case where this would fail is that of a sequence of M tosses of a biased coin, where the experimenter only reports that initial sequence maximising the excess of heads over tails. In Dawid and Dickey (1977), it is shown how to adjust the 'face-value likelihood function' to account for this biased reporting process. I wonder if there might be any parallel to this in the betting context?

I have been greatly stimulated by Glenn Shafer's highly original ideas over many decades, and am grateful for this opportunity to express my appreciation by proposing a very warm vote of thanks.

REFERENCES

Dawid, A.P. & Dickey, J.M. (1977) Likelihood and Bayesian inference from selectively reported data. *Journal of the American Statistical Association*, 72, 845–850.

Duke, A. (2018) *Thinking In Bets: Making Smarter Decisions When You Don't Have All The Facts*. New York: Portfolio Penguin.

Fraser, D.A.S. (1968) *The Structure Of Inference*. Hoboken, NJ: Wiley.

How to cite this article: Philip A. Dawid. Proposer of the vote of thanks to Glenn Shafer and contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication'. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12648

DOI: 10.1111/rssa.12649

Seconder of the vote of thanks to Glenn Shafer and contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication'

Frank P. A. Coolen
Durham University, Durham, UK

Correspondence: Frank P.A. Coolen, Durham University, Durham, UK.
Email: frank.coolen@durham.ac.uk

Professor Shafer's paper, as so much of his work, has taught me interesting new aspects of statistical inference with substantial historical context. Testing a probability distribution by betting is simple and powerful, and the betting interpretation is natural for sequential testing of hypotheses. The fact that the betting score is a likelihood ratio and the implied targets are very interesting, as is the further justification for the Neyman–Pearson theory in case of a one-off hypothesis test.

My main query is about application. In practice, there is often no clear random phenomenon Y of interest, but, for example a wish to show that one method is better than another. Translating the practical question to a statistical scenario is important, the choice of Y can greatly influence the apparent conclusion. One typically designs an experiment resulting in a sample and summaries of the sample observations may be of interest. Suppose one observes Y_i, $i=1,\ldots,n$, and claims that these are observations of independent random quantities from distribution P. Let an alternative distribution Q be proposed, with the same mean as P but larger variance. One could use a test on the Y_i's or the mean $\overline{Y}$. If the data mean is close to the mean of P and Q, but the data variance is large, then the bet on $\overline{Y}$ is likely to support P while the bet on the Y_i's is likely to support Q. An experienced statistician will understand that these two tests are different (similar to Example 4 in the paper), but others may be confused. It raises the question who decides on the choice of Y. Is it possible, perhaps by using the implied targets, to find a statistic Y such that P and Q can be distinguished in some optimal manner?

I note that Y must be fully known in order to apply the theory, which means that the design of the experiment or the way data are collected must be known, how otherwise can one assign a meaningful P and S (or Q)? This is important not only for small-scale experiments, but also in applications with very large amounts of data, and I fear it is easily overlooked if statistical methods are applied without care. Elicitation of probability distributions based on expert judgements is difficult, but the fact that Y needs to be observable (or a function of observables) will be helpful. It will be important to see if S can be elicited, or if focus on Q works better to formulate the alternative to P.

Elicitation reminds me of my first statistics application, which considered reliability of heat exchangers (Coolen et al., 1992). Inspections planning required expert judgements as inputs, and the available experts' opinions varied substantially. I used linear opinion pooling in a Bayesian setting, learning about the levels of expertise when data became available. Starting off with equal weights

for the experts, these were updated to become proportional to the probabilities of the observed data according to the experts (the denominator in Bayes' theorem), which has similarities with the probabilities of the data being used to distinguish between the claimed hypothesis (P) and the alternative (Q) in testing by betting. These weights had the property that it did not matter for the overall inference, based on the linearly pooled expert opinions, whether the opinions were first pooled and then updated or vice versa. The decision maker could interpret these weights as proportions of overall budget to assign to the experts. The work did not consider betting scores or hypothesis testing, but I believe there are links to the use of expert judgement in decision problems that could be explored.

Professor Shafer begins the paper with the statement that the p-value is too complicated for effective communication to a wide audience. This has received increased attention in recent years as part of the discussion about repeatability and reproducibility of experiments (Atmanspacher & Maasen, 2016; Goodman, 1992; Senn, 2002). In recent work, we have considered reproducibility of hypothesis tests from non-parametric predictive inference perspective (Augustin & Coolen, 2004; Coolen & Bin Himd, 2014; Coolen & Marques, 2020), which shows that the implicit statistical reproducibility is often poor, in particular for multi-group tests with one-sided alternatives. This issue will not be resolved using testing by betting, as it is a direct consequence of the dichotomous nature of such tests. Of course, one would like to overcome this problem by gathering more data when needed, that is if the test criterion is close to the decision borderline, the newly proposed methods may be useful here.

I am not sure if testing by betting will make statistics easier, I believe that statistics requires expertise as it is a very challenging topic bridging pure mathematics and real world applications in many fields, as nicely discussed by Hampel (Hampel, 1998), who also set out to develop an objective theory of 'successful betting' (Hampel, 2001). In this paper, Professor Shafer has presented a useful new tool for expert mathematical statisticians, which requires further development for practical implementation. It gives me great pleasure to second the vote of thanks.

REFERENCES

Augustin, T. & Coolen, F.P.A. (2004) Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124, 251–272.

Atmanspacher, H. & Maasen, S. (2016) *Reproducibility: Principles, problems, practices, and prospects*. Hoboken: Wiley.

Coolen, F.P.A. & Bin Himd, S. (2014) Nonparametric predictive inference for reproducibility of basic nonparametric tests. *Journal of Statistical Theory and Practice*, 8, 591–618.

Coolen, F.P.A. & Marques, F.J. (2020) Nonparametric predictive inference for test reproducibility by sampling future data orderings. *Journal of Statistical Theory and Practice*, 14, 2.

Coolen, F.P.A., Mertens, P.R. & Newby, M.J. (1992) A Bayes-competing risk model for the use of expert judgment in reliability estimation. *Reliability Engineering and System Safety*, 35, 23–30.

Goodman, S.N. (1992) A comment on replication, p-values and evidence. *Statistics in Medicine*, 11, 875–879.

Hampel, F. (1998) Is statistics too difficult? *The Canadian Journal of Statistics*, 26, 497–513.

Hampel, F. (2001) An outline of a unifying statistical theory. In: *Proceedings of the 2nd international symposium on imprecise probabilities and their applications, Ithaca, New York, 2001*.

Senn, S. (2002) Comment on 'A comment on replication, p-value and evidence', by S.N. Goodman (Letter to the editor). *Statistics in Medicine*, 21, 2437–2444 (with reply by S.N. Goodman, pp. 2445–2447).

How to cite this article: Coolen FP. Seconder of the vote of thanks to Glenn Shafer and contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication'. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12649

The vote of thanks was passed by acclamation.

DOI: 10.1111/rssa.12650

Harry Crane's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Harry Crane
Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ, USA

Correspondence: Harry Crane, Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ, USA.
Email: hcrane@stat.rutgers.edu

Betting is central to the history of probability and the way probability is intuitively understood, making it both natural to link betting to statistical analysis and curious that this connection is absent from conventional statistical thinking. Betting appears in Bayesian foundations, but as a philosophical relic rather than a substantive component of the framework. When the devout Bayesian talks about betting, he only cares that the bettor's probabilities make sense (are coherent), not whether they make money. Shafer's view is more palatable for science, where personal opinions should take a back seat to objective reality—though sadly this is not always the case (Crane et al., 2020).

Shafer's proposal may improve how statistical work is communicated, but it stops short of what is needed to resolve systemic statistical abuse. In advocating his theory, Shafer writes, 'I need not risk a lot of money. [...] I am betting merely to make a point'. But what is the point of a fictitious bet?

In gambling parlance, a *freeroll* is a bet that can be won but not lost. In scientific work, the *Freeroll Effect* occurs when scientists incur minimal personal risk in exchange for broad societal impact. Scientists are rewarded for publishing their research in high-impact journals while society bears the risk of inaccurately reported findings and their potentially dire consequences. The replication crisis, misinformed COVID-19 response, and muddled climate policies are all consequences of the Freeroll Effect (Crane, 2020).

So, yes, the amount risked does matter. As any gambler knows, there is a big difference between betting a penny and a thousand dollars. It is not 'irrational', as a Bayesian might claim. It is common sense. We should hope that scientists exercise this same sense before publishing research that burdens society with substantial risk.

Fortunately, mathematical probability has a built-in property to achieve this objective, called the *Fundamental Principle of Probability* (FPP) (Crane, 2018). Under the FPP, statistical claims are meaningless, and should be disregarded, unless the statistician faces real-world consequences for being wrong. Some critics of the FPP offer the jejune moral objection that gambling is lowbrow, having no place in science. Apart from their deep misunderstanding of risk and its central role in probability, such critics exhibit disregard for the serious practical problems that can be resolved by appealing to risk in statistical practice.

So I advocate to take Shafer's proposal even more seriously than he suggests, by restoring fundamental principles of probability and risk to statistical work, not simply using a different language.

REFERENCES

Crane, H. (2018) The fundamental principle of probability. Researchers.One. https://www.researchers.one/article/2018-08-16

Crane, H. (2020) Naive probabilism. Researchers. One. https://www.researchers.one/article/2018-03-9

Crane, H., Guinness, J. & Martin, R. (2020) Comment on the proposal to rename the R.A. Fisher lecture. Researchers. One. https://www.researchers.one/article/2020-06-11

How to cite this article: Crane H. Harry Crane's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12650

DOI: 10.1111/rssa.12651

Barbara Osimani's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Barbara Osimani
UNIVPM, Ancona, Italy

Correspondence: Barbara Osimani, UNIVPM, Ancona, Italy.
Email: b.osimani@univpm.it

With respect to standard frequentist approaches to hypothesis testing, the testing by betting (TbB) approach has the considerable advantage of retaining much more of the information provided by the evidence, in that the betting score is given by the distance between two probability distributions (preferably measured in terms of Kullback-Leibler divergence) - in contrast to traditional "pointwise" measures of evidence. Manifestly, such measure is much more informative about the impact of the observed evidence than *p*-values or Neyman-Pearson significance thresholds.

This explains the intuition that the TbB approach more reliably reflects the evidential impact of observations in the various experimental settings presented in the examples, than *p*-values and Neyman–Pearson decisions do. The approach also allows one to test hypotheses opportunistically, that is, without a predefined plan over an established sequence of variables.

All the above advantages, however, do not give any special standing to the TbB approach over Bayesian instruments of scientific inference. Since I do not think that the TbB approach per se is easier to understand or to communicate than standard approaches to statistical inference either—be they frequentist or Bayesian style—one needs stronger reasons to prefer TbB to Bayesianism.

Such reasons may be given by its potentially higher flexibility in accommodating the joint impact of multiple evidence about various dimensions of the inferential set-up, on the target hypothesis. That is, not only direct evidence about such hypothesis, but also (educated) estimates about the validity of the statistical model itself, which may moderate the impact of the former.

The combination of bets over these first and secondary order hypotheses may represent the joint inferential import of such dimensions of evidence. This would align with recent appeals to more comprehensive approaches to evidence in response to the so called 'reproducibility crisis' (see Gelman, 2015; Gelman & Carlin, 2017; Osimani, 2020).

In emphasizing the role of noise at the root of such crisis, Gelman (2015) reminds us that 'in Bayesian inference data may be combined with prior information, e.g. the prior expectation that newly studied effects tend to be small, which leads us to downwardly adjust large estimated effects in light of the high probability that they could be coming largely from noise'. Although he does not advance a full-fledged hierarchical Bayes solution, his advices may foster this direction as well.

Is there any opportunity to accommodate Gelman's proposal in the TbB approach? More specifically, is it possible to combine intuitions about the target hypothesis itself and about the parameters of the statistical model with which we test it? Also, can sceptic make money by betting against too precise hypotheses?

REFERENCES

Gelman, A. (2015) Working through some issues. *Significance*, 12, 33–35. https://doi.org/10.1111/j.1740-9713.2015.00828.x.

Gelman, A. & Carlin, J. (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. https://doi.org/10.1080/01621459.2017.1311263.

Osimani, B. (2020) Social games and epistemic losses: Reliability and higher order evidence in medicine and pharmacology. In: Osimani, B. and La Caze, A. (Eds.) *Uncertainty in Pharmacology: Epistemology, methods and decisions*. Berlin: Springer, Boston Series in Philosophy of Science, pp. 345–372.

How to cite this article: Osimani B. Barbara Osimani's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12651

DOI: 10.1111/rssa.12652

Aaditya Ramdas's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Aaditya Ramdas
Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

Correspondence: Aaditya Ramdas, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.
Email: aramdas@cmu.edu

I congratulate Shafer (2020) on his eloquent arguments for the adoption of betting as a language for communicating uncertainty. While I find myself largely in agreement with his perspectives, I hope to complement his ideas by presenting here what I believe are some orthogonal, 'purely

statistical' reasons to embrace betting scores and their measure-theoretic cousins: e-values and martingales.

Let the null be given by a composite set of distributions $\mathcal{P}$; an e-variable E is a non-negative random variable such that $\mathbb{E}_P[E] \leq 1$ for any $P \in \mathcal{P}$; the instantiation of an e-variable is an e-value. A process $(M_t)_{t\in\mathbb{N}}$ is a non-negative $\mathcal{P}$-martingale ($\mathcal{P}$-NM) if it is a non-negative martingale under every $P \in \mathcal{P}$.

Reinterpreting Wald (2004) in terms of betting. As an historical aside, one can reinterpret an approach by Wald in terms of betting. In his 1940s textbook (Wald, 2004), Eq. (10:10) describes a method for testing a point null $X \sim f_{\theta_0}$ for some parametric class of densities $(f_\theta)_{\theta\in\Theta}$. There, Wald suggests using $S_i := \frac{f_{\widehat{\theta}_{i-1}}}{f_{\theta_0}}$ to 'bet' on X_i, where $\widehat{\theta}_{i-1}$ is any estimator of θ from $X_1, \ldots, X_{i-1}$. The bet S_i is clearly predictable and has unit expectation under the null. One rejects when $M_t \geq 1/\alpha$ for $M_t = \prod_{i=1}^{t} S_i(X_i)$; type I error is controlled by Ville's inequality applied to the P_{θ_0}-NM capital process $(M_t)_{t\in\mathbb{N}}$.

NMs are likelihood ratios, capital processes and admissible e-values (Ramdas et al., 2020; Shafer et al., 2011). Beyond the singleton parametric setting above, $\mathcal{P}$-NMs can be constructed for several rich non-parametric classes $\mathcal{P}$ (Howard et al., 2020). Not only is a likelihood ratio dQ/dP a pointwise P-NM, but it is known that every composite $\mathcal{P}$-NM can be represented as a likelihood ratio dQ/dP for every $P \in \mathcal{P}$ (Ramdas et al., 2020; Shafer et al., 2011). (Here Q plays the role of Shafer's implied alternative for a bet against P, and the $\mathcal{P}$-NM is thus a capital process.) These $\mathcal{P}$-NMs can also be used to construct non-parametrically valid 'confidence sequences' (Howard et al., 2020), measure-theoretic analogues of Shafer's warranty sets. In fact, composite $\mathcal{P}$-NMs (coupled with the optional stopping theorem) yield admissible 'safe' e-values in a concrete sense (Grünwald et al., 2019; Ramdas et al., 2020). However, not every admissible e-value needs to be a composite $\mathcal{P}$-NM, but must be the infimum of pointwise P-NMs (Ramdas et al., 2020).

Dimension-agnostic, universal inference (Wasserman et al., 2020). In non-sequential settings, a classical approach for constructing tests and confidence sets is to use Wilks' asymptotics for the likelihood ratio. However, in high-dimensional settings, when the dimension d and sample size n approach infinity together, the limiting distribution F of the likelihood ratio can differ from the low-dimensional setting. In order to derive F when $n = 1000$, $d = 200$, should we assume that $n \gg d$ or $d/n \to \kappa \in (0,1)$? Furthermore, even low-dimensional asymptotics are often intractable for irregular composite nulls. In contrast, universal inference (Wasserman et al., 2020) is non-asymptotically valid for any d,n under no regularity conditions; the relevant point is that the central object in universal inference is an e-value called the split/crossfit likelihood ratio.

Multiple testing under dependence (Wang & Ramdas, 2020). All of the above points are related to the study of e-values for a single hypothesis. It is worth noting that e-values also have some significant advantages in multiple testing (Vovk & Wang, 2020). For instance, the e-BH procedure (Wang & Ramdas, 2020), which is the natural e-value analogue of the Benjamini–Hochberg (BH) procedure, controls the false discovery rate *under arbitrary dependence with no correction factor*. Stated differently, if we use p-values derived from e-values as $p_i := 1/e_i$ (say, using e-values from universal inference, or stopped supermartingales), then the BH procedure is automatically robust to arbitrary dependence with no correction.

In summary, (composite, non-parametric) betting-scores, e-values and martingales provide several advantages and new avenues for inference that are complementary to classical approaches. These instruments are worthy of deeper study irrespective of philosophical stances on their broader roles.

REFERENCES

Grünwald, P., de Heide, R. & Koolen, W. (2019) Safe testing. *arXiv:1906.07801*.

Howard, S.R., Ramdas, A., McAuliffe, J. & Sekhon, J. (2020) Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17, 257–317.

Howard, S.R., Ramdas, A., McAuliffe, J. & Sekhon, J. (2020) Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, forthcoming.

Ramdas, A., Ruf, J., Larsson, M. & Koolen, W. (2020) Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*.

Shafer, G. (2020) The language of betting as a strategy for statistical and scientific communication (with discussion). *Journal of the Royal Statistical Society, Series A*.

Shafer, G., Shen, A., Vereshchagin, N. & Vovk, V. (2011) Test martingales, Bayes factors and *p*-values. *Statistical Science*, 26(1), 84–101.

Vovk, V. & Wang, R. (2020) E-values: Calibration, combination, and applications. *The Annals of Statistics*, forthcoming.

Wald, A. (2004) *Sequential analysis*. Chelmsford, MA: Courier Corporation.

Wang, R. & Ramdas, A. (2020) False discovery rate control with e-values. *arXiv preprint arXiv:2009.02824*.

Wasserman, L., Ramdas, A. & Balakrishnan, S. (2020) Universal inference. In: *Proceedings of the national academy of sciences*. ISSN 0027-8424.

How to cite this article: Ramdas A. Aaditya Ramdas's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12652

DOI: 10.1111/rssa.12653

Peter D. Grünwald's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Peter D. Grünwald
CWI and Leiden University, Amsterdam, The Netherlands

Correspondence: Peter D. Grünwald, CWI and Leiden University, Amsterdam, The Netherlands.
Email: pdg@cwi.nl

Evidence can be effectively summarized and reported in the form of *betting scores*, perhaps more commonly known these days as *E-values* (Grünwald et al., 2019; Vovk & Wang, 2019; Wang & Ramdas, 2020). Professor Shafer is to be congratulated for an exceedingly clear exposition of this idea!

Composite Nulls Professor Shafer defines the *betting score against composite null* underneath Protocol 5 in Section 4.2 of his paper. Shafer's definition superficially looks quite different from the one in Grünwald et al., 2019, a paper focusing on composite nulls. We now show that, fortunately, the two definitions are equivalent after all. For simplicity, we modify Shafer's setting to deal with a one-shot bet for a random sample $(Y_1, \ldots, Y_n)$ of fixed size n for which we fix some parameterized model $\{P_\theta : \theta \in \Theta\}$. Take any $\theta_0 \in \Theta$. Both papers define a betting score or E-variable against simple null $\{\theta_0\}$ as a non-negative random variable S such that $\mathbf{E}_{P_{\theta_0}}[S] \leq 1$.

Now take as null the full, composite $\Theta_0 := \Theta$. Let $\mathcal{S}(\Theta_0) := \{S_\theta : \theta \in \Theta_0\}$ be an arbitrary collection of betting scores, that is for all $\theta \in \Theta_0$, S_θ is a betting score against $\{\theta\}$. Shafer defines the betting score against Θ_0, defined relative to $\mathcal{S}_{\Theta_0}$, as $S_{\mathcal{S}(\Theta_0)} := inf_{\theta \in \Theta_0} S_\theta$. Thus, any collection of valid betting scores against each individual $\theta \in \Theta_0$ gives a valid betting score against Θ_0. The betting interpretation is straightforward: if, under each $\theta \in \Theta_0$, we do not expect to gain any money under pay-offs S_θ, we certainly do not expect to gain any money, no matter from what $\theta \in \Theta_0$ data are sampled, if we get paid the minimum S_θ.

In contrast, Grünwald et al. (2019) define an E-variable against Θ_0 to be any non-negative random variable S satisfying that

$$\textit{for all } \theta \in \Theta_0, \quad \mathbf{E}_{P_\theta}[S] \leq 1. \tag{1}$$

To see that the definitions coincide, note that for any $\mathcal{S}(\Theta_0)$, Shafer's $S_{\mathcal{S}(\Theta_0)}$ satisfies

$$\text{for all } \theta \in \Theta_0 : \mathbf{E}_{P_\theta}[S_{\mathcal{S}(\Theta_0)}] = \mathbf{E}_{P_\theta}[\inf_{\theta' \in \Theta_0} S_{\theta'}] \leq \mathbf{E}_{P_\theta}[S_\theta] \leq 1,$$

so any Shafer-betting score against composite Θ_0 is also an E-variable relative to Θ_0 in our sense. Conversely, let S be any E-variable relative to composite Θ_0 in our sense. Define $\mathcal{S}(\Theta_0) := \{S_\theta : \theta \in \Theta_0\}$ with $S_\theta := S$ and hence by construction $\mathbf{E}_{P_\theta}[S_\theta] \leq 1$ for all $\theta \in \Theta_0$, so $\mathcal{S}(\Theta_0)$ is a valid collection of betting scores against each θ. But also $S_{\mathcal{S}(\Theta_0)} = inf_{\theta \in \Theta_0} S_\theta = S$, so our E-variable S is also a valid betting score against Θ_0 in Shafer's sense.

Betting Scores and Bayes Factors Equation (1) also illustrates that, in general, Bayes factors for composite Θ_0 are *not* betting scores. Any Bayes factor can be written as $B = \int p_\theta(Y^n) dW_1(\theta) / \int p_\theta(Y^n) dW_0(\theta)$ with W_1, W_0 priors over Θ_1 and Θ_0. It satisfies (Grünwald et al., 2019)

$$\mathbf{E}_{\theta \sim W_0} \mathbf{E}_{P_\theta}[B] = 1.$$

Betting scores require the much stronger property (1) that the no-expected-gain property holds for all $\theta \in \Theta_0$, and not just in expectation over a given prior. For many popular Bayes factors, Equation (1) does not hold (Grünwald et al., 2019). On the other hand, using a minimax theorem from Grünwald and Dawid (2004), they show how to construct, for arbitrary W_1, a special 'matching' prior W_0 such that the corresponding Bayes factor B does satisfy Equation (1). For some models, this construction produces well-known priors, such as the right Haar prior on a nuisance variance parameter; for other models, the resulting priors have not been considered before.

REFERENCES

Grünwald, P.D. & Dawid, A.P. (2004) Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32(4), 1367–1433.

Grünwald, P., de Heide, R. & Koolen, W. (2019) Safe testing. *arXiv preprint arXiv:1906.07801*.

Vovk, V. & Wang, R. (2019) Combining e-values and p-values. *Available at SSRN*.

Wang, R. & Ramdas, A. (2020) False discovery rate control with e-values. *arXiv preprint arXiv:2009.02824*.

How to cite this article: Grünwald PD. Peter D. Grünwald's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. Grünwald. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12653

DOI: 10.1111/rssa.12654

Xiao-Li Meng's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Xiao-Li Meng
Harvard University, Cambridge, MA, USA

Correspondence: Xiao-Li Meng, Harvard University, Cambridge, MA, USA.
Email: meng@stat.harvard.edu

Once in a while, an RSS' reading article provides an undulated/undiluted intoxication that only a few lucky wine connoisseurs have experienced. A well sought-after Grand Cru was aired with much anticipation. It delivered, but in unexpected ways. The betting interpretation of probability is as old as probability itself. Using likelihood ratio as a metric for evidence also has a long history, with the virtue of being relevant (because it avoids hypothetical replications) *and* robust (to prior specification). As discussed in Liu and Meng (2016), the price one pays for this seemingly impossible combination is the lack of interpretability because '"likelihood units" do not have an operational/real life meaning'. Surprisingly, then, Professor Glenn's betting score endows the likelihood ratio exactly such a meaning, at least for discrete data. However, its effectiveness with continuous data might need some further testing, since a ratio of densities is harder to explain (to general public) than a ratio of probabilities.

Nevertheless, mapping a given betting score $S(y)$ into a model for an alternative hypothesis has a similar fruitful flavour as reverse engineering a test statistic T into a local alternative model (e.g. Kong & Cox, 1997). The latter was done via exponential tilting $P_\theta(y) = P_0(y)\exp\{\theta T(y)\}/c_\theta$, where $\theta = 0$ indexes the null hypothesis. This construction directly links T with the *log-likelihood* ratio, and hence, it does not prompt the kind of motivation in Section 2.2 going from S to ln S. The log-scale also motivated Nicolae et al. (2008) to invoke the 'extended EM identity'

$$\mathrm{E}\left[\log\frac{P_{\theta_0}(Y_c)}{P_{\theta_1}(Y_c)}\middle| Y_o,\theta\right] = \log\frac{P_{\theta_0}(Y_o)}{P_{\theta_1}(Y_o)} + \mathrm{E}\left[\log\frac{P_{\theta_0}(Y_c\,|\,Y_o)}{P_{\theta_1}(Y_c\,|\,Y_o)}\middle| Y_o,\theta\right] \tag{1}$$

for measuring information in $T(y)$, when we move from the observed $y = Y_o$ to the ideal complete data $y = Y_c$, a design consideration for determining the cost benefit of collecting more data. By choosing $\theta = \theta_0$, we obtain the Gibbs's inequality used in Section 2.2, since the second term on the right-hand side of Equation (1) then becomes the K-L divergence between $P_{\theta_0}(Y_c\,|\,Y_o)$ and $P_{\theta_1}(Y_c\,|\,Y_o)$.

It is also interesting to contrast Equation (1) with the martingale property of likelihood ratio discussed in the article's Appendix, which renders

$$\mathrm{E}\left[\frac{P_{\theta_0}(Y_c)}{P_{\theta_1}(Y_c)}\middle| Y_o,\theta_1\right] = \frac{P_{\theta_0}(Y_o)}{P_{\theta_1}(Y_o)}. \tag{2}$$

That is, $\mathrm{E}\left[S(Y_c) \mid Y_o, \theta_1\right] = S(Y_o)$, indicating the coherence of the betting score interpretation with respect to (our normal) expectation. A third relevant identity, a Bayesian counterpart to Equation (2), is obtained when both P and Q are indexed by $\theta \in \Theta$, for which we have a prior π such that both posteriors $P(\theta|y)$ and $Q(\theta|y)$ are proper. If our payoff for betting against a $P(y|\theta)$ is $S(y;\theta) = Q(y|\theta)/P(y|\theta)$ when θ is known,[1] then we would expect the payoff is $\mathrm{E}[S(y;\theta)|y]$ when θ is unknown. However, as given in Section 6.5 of Nicolae et al. (2008),

$$\mathrm{E}\left[S(y;\theta) \mid y\right] = \int \frac{Q(y \mid \theta)}{P(y \mid \theta)} P(\theta \mid y)\mu(d\theta) = \int \frac{Q(y \mid \theta)\pi(\theta)}{P(y)}\mu(d\theta) = \frac{Q(y)}{P(y)}, \tag{3}$$

which is just the Bayes factor for comparing Q and P. Identity (3) seems to further justify considering a likelihood ratio as a betting score (again, most directly when y is discrete).

REFERENCES

Kong, A. & Cox, N.J. (1997) Allele-sharing models: Lod scores and accurate linkage tests. *The American Journal of Human Genetics*, 61(5), 1179–1188.

Liu, K. & Meng, X.-L. (2016) There is individualized treatment. Why not individualized inference? *Annual Review of Statistics and its Application*, 3(1), 79–111.

Nicolae, D.L., Meng, X.-L. & Kong, A. (2008) Quantifying the fraction of missing information for hypothesis testing in statistical and genetic studies (with discussions). *Statistical Science*, 23(3), 287–331. Available from: https://doi.org/10.1214/08-STS244REJ.

How to cite this article: Meng XL. Xiao-Li Meng's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12654

[1]A Caveat: It requires $S(y;\theta)$ to be first-order ancillary: $\mathrm{E}_{P_\theta}[S(Y;\theta)] = 1$ for all $\theta \in \Theta$.

DOI: 10.1111/rssa.12655

Arthur Paul Pedersen's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Arthur Paul Pedersen
Department of Computer Science, City College of New York, The City University of New York (CUNY), New York, NY, USA

Correspondence: Arthur Paul Pedersen, Department of Computer Science, City College of New York, The City University of New York (CUNY), New York, USA.
Email: apedersen@cs.ccny.cuny.edu

I would like to add my congratulations to the author for his important paper contribution for reading. The author's wager taken before this audience with such brilliant candour does not go unnoticed—this deliberate on-themed-ness is to be imputed to the author's total wickedness.

The many substantive issues raised during this discussion meeting fail to do justice to the breadth and depth of questions calling for careful attention. Here I limit myself to two pint-sized questions, clumsily raised, about the author's paper.

The author repeatedly insists that testing by betting need not risk a lot of cash; sufficiently small amounts of cash, or even play cash, is enough for betting to perform its function, to make a point. Testing by betting appeals neither to utility evaluations nor Bayesian reasoning—it is, the author contends, 'not a chapter in decision theory.'

Bruno de Finetti (1974), who uses cash payoffs through bets, contracts, or scoring rules in probability elicitation, likewise argues that non-linearities in utility evaluations can be ignored for practical purposes on condition that cash payoffs be sufficiently small, an assumption de Finetti coins the 'Hypothesis of Rigidity.' De Finetti, albeit superhuman, is mistaken in his reasoning. The Hypothesis of Rigidity fails to rule out the impact of utility evaluations in probability elicitation should, say, a probability assessor have skin in the game on stochastic events for which probabilities are to be elicited.

On this thread, Kadane and Winkler (1987) show that while the Hypothesis of Rigidity alone thus fails to eliminate the impact of utility evaluations in probability elicitation, it decidedly succeeds in doing so in the presence of the additional requirement that the probability assessor's fortune be independent of events for which probabilities are being elicited, discounting payoffs borne out of elicitation (cf. Kadane & Winkler, 1988).

For my first question, I respectfully ask our distinguished author to please explain whether, and perhaps the extent to which, the foregoing considerations have bearing on developments from the paper under discussion.

In part due to space requirements, the second question I curb on the way out: I respectfully ask the author to explain whether jettisoning his paper's prominent claims about simplicity would negatively affect the significance of what is already a first-rate paper contribution.

REFERENCES

de Finetti, B. (1974) *Theory of probability*, volume I, 1990 edition. Hoboken, NJ: Wiley.

Kadane, J.B. & Winkler, R.L. (1987) De finetti's methods of elicitation. In: R. Viertl (Ed.) *Probability and Bayesian statistics*. Boston, MA: Springer, pp. 279–284.

Kadane, J.B. & Winkler, R.L. (1988) Separating probability elicitation from utilities. *Journal of the American Statistical Association*, 83(402), 357–363.

How to cite this article: Pedersen AP. Arthur Paul Pedersen's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12655

DOI: 10.1111/rssa.12656

Vladimir Vovk's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Vladimir Vovk
Royal Holloway, University of London, Egham, Surrey, UK

Correspondence: Vladimir Vovk, Royal Holloway, University of London, Egham, Surrey, UK.
Email: v.vovk@rhul.ac.uk

Glenn Shafer's paper is a powerful appeal for a wider use of betting ideas and intuitions in statistics. He admits that *p*-values will never be completely replaced by betting scores, and I discuss it further in Vovk (2020a) (Appendix A) (one of the two online appendices that I have prepared to meet the word limit). Both *p*-values and betting scores generalise Cournot's principle (Shafer, 2007), but they do it in their different ways, and both ways are interesting and valuable.

Other authors have referred to betting scores as Bayes factors (Shafer et al., 2011) and *e*-values (Grünwald et al., 2020; Vovk & Wang, 2020). For simple null hypotheses, betting scores and Bayes factors indeed essentially coincide (Grünwald et al., 2020, Section 1, interpretation 3), but for composite null hypotheses they are different notions, and using 'Bayes factor' to mean 'betting score' is utterly confusing to Bayesians (Robert, 2011). However, the Bayesian connection still allows us to apply Jeffreys's (1961, Appendix B) rule of thumb to betting scores, namely, a *p*-value of 5% is roughly equivalent to a betting score of $10^{1/2}$, and a *p*-value of 1% to a betting score of 10. This agrees beautifully with Shafer's rule (6), which gives, to two decimal places:

1. for p=5%, 3.47 instead of Jeffreys's 3.16 (slight overshoot);
2. for p=1%, 9 instead of Jeffreys's 10 (slight undershoot).

The term '*e*-values' emphasises the fundamental role of expectation in the definition of betting scores (somewhat similar to the role of probability in the definition of *p*-values). It appears that the natural habitat for 'betting scores' is game-theoretic while for '*e*-values' it is measure-theoretic (Shafer, 2020); therefore, I will say '*e*-values' in the online appendices (Vovk, 2020a,b), which are based on measure-theoretic probability.

In the second online appendix (Vovk, 2020b), I give a new example showing that betting scores are not just about communication; they may allow us to solve real statistical and scientific problems (more examples are given in the comment by my co-author Ruodu Wang). David Cox (1975) discovered that splitting data at random not only allows flexible testing of statistical hypotheses but also achieves high efficiency. A serious objection to the method is that different people analysing the same data may

get very different answers (thus violating 'inferential reproducibility', Goodman et al., 2016; Held & Schwab, 2020). Using *e*-values instead of *p*-values remedies the situation.

REFERENCES

Cox, D.R. (1975) A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62, 441–444.

Goodman, S.N., Fanelli, D. & Ioannidis, J.P.A. (2016) What does research reproducibility mean? *Science Translational Medicine*, 8, 341ps12.

Grünwald, P., de Heide, R. & Koolen, W.M. (2020) Safe testing. *Technical Report arXiv:1906.07801 [math.ST]*, arXiv.org e-Print archive.

Held, L. & Schwab, S. (2020) Improving the reproducibility of science. *Significance*, 17(1), 10–11.

Jeffreys, H. (1961) *Theory of probability*, 3rd edn. Oxford: Oxford University Press.

Robert, C.P. (2011) Bayes factors and martingales. Entry in blog "Xi'an's Og" for August 11.

Shafer, G. (2007) From Cournot's principle to market efficiency. In: J.-P. Touffut (Ed.) *Augustin Cournot: Modelling economics*, chap. 4. Cheltenham: Edward Elgar.

Shafer, G. (2020) Personal communication. May 8.

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

Vovk, V. (2020a) Comment on Glenn Shafer's "Testing by betting". *Technical Report arXiv:2009.08933 [stat.ME]*, arXiv.org e-Print archive. Appendix A: Cournot's principle, p-values, and e-values.

Vovk, V. (2020b) A note on data splitting with e-values: Online appendix to my comment on Glenn Shafer's "Testing by betting". *Technical Report arXiv:2008.11474 [stat.ME]*, arXiv.org e-Print archive.

Vovk, V. & Wang, R. (2020) E-values: Calibration, combination, and applications. *Technical Report arXiv:1912.06116 [math.ST]*, arXiv.org e-Print archive. To appear in the *Annals of Statistics*.

How to cite this article: Vovk V. Vladimir Vovk's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12656

DOI: 10.1111/rssa.12657

Christian Hennig's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Christian Hennig
University of Bologna, Bologna, BO, Italy

Correspondence: Christian Hennig, University of Bologna, Bologna BO, Italy.
Email: christian.hennig@unibo.it

I think that Shafer's paper, despite some interesting ideas, fails at its core aim of enabling a better communication of statistical evidence.

1. Reasons why *p*-values are often misinterpreted have been discussed elsewhere (e.g. Mayo, 2018). Many journals made reaching significance a condition for publication, incentivising misuse. I believe that statistical inference, however done, is inherently difficult, and issues are to be expected. There is no agreement on the foundations of probability. People want certainty and we do not deliver it. People want statements about truth, but our models are never 'true'.
2. If a test of $H_0: P = P_0$ is based on a statistic T with rejection region, say, $\{T > c\}$, then H_0 is tested against the implicit alternative of *any* Q with $Q\{T > c\} > P_0\{T > c\}$. There is no reason, when interpreting a test result, to commit to any more precise specification of Q. All examples in Section 2.4 rely on unrealistically precisely specified alternatives. I like the definition of the 'implied alternative' as a tool to understand tests better, but it is a sophisticated concept, and rather unhelpful for simple communication of test results. It may encourage wrongly inferring specific alternatives from tests.
3. Betting is by no means a straightforward and unproblematic concept. Shafer's communication seems to imply that scientists should be prepared to bet, meaning that they should favour outcomes that lead to them winning. This is not an optimal attitude for science. There is payoff in case of rejection of the null, and Shafer implies that the scientist hopes to achieve such a payoff (Shafer aims at improving communication, so his communication is of key importance). This seems to take the incentive of journals for finding significance for granted, which is one of the major issues with classical tests. (Later setups in which the scientist does not bet herself look rather convoluted.)
4. Not everybody, including scientists, is comfortable or familiar with betting. Much work (e.g. Tversky & Kahnemann, 1974) suggests that real betting behaviour is often not rational. Betting behaviour and attitudes to betting differ between different parts of society, for example between men and women (Wong et al., 2013); the quote by Kelly in the paper takes for granted that the gambler is a man, and his wife's role is very different. There is a substantial amount of problem gambling. All this leaves me very worried about giving the betting metaphor a major role in statistics communication.

REFERENCES

Mayo, D.G. (2018) *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

Tversky, A. & Kahnemann, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.

Wong, G., Zane, N., Saw, A., & Chan, A.K. (2013) Examining gender differences for gambling engagement and gambling problems among emerging adults. *Journal of Gambling Studies*, 29, 171–189.

How to cite this article: Hennig C. Christian Hennig's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12657

DOI: 10.1111/rssa.12671

A distillation of the live chat during the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Paul A. Smith

Correspondence: Paul A. Smith, Discussion Papers Editor
Email: p.a.smith@soton.ac.uk

Professor Shafer's paper generated a lot of interest in the online chat, and this contribution is a selection distilled from there. Several discussants wanted a simple worked example to show how they could use the betting approach in a practical situation—say, as a way of explaining an outcome to a client. Since the vast majority of published tests are not done by professional statisticians, the betting approach will only work if they can see how it is used. Grünwald et al. (2019, section 5.2) was suggested to show a simple *t*-test example, and there is a vignette at https://safestatistics.com/safe-testing-vignette/.

Professor Shafer (and other authors of different approaches) has claimed that alternatives to *p*-values are more easily interpreted or less prone to misinterpretation, but there was a call for experiments to be undertaken to provide evidence for whether this is true in practice (DP Editor's note: which approach would you use to test the outcome of the experiment?). This kind of evidence will be needed to persuade journals to accept or recommend the betting approach. Until journals actively encourage these methods, nothing is likely to change.

Does betting (or hypothesis testing) answer the question users are asking? One thought was that confidence intervals (or distributions) correspond better with user requirements, while another suggested calculating the probability of the null hypothesis being true given the observed critical value (or bet outcome)—as this is the interpretation some unsophisticated users give (incorrectly) to standard tests.

If we accept betting, how much should the return be? We have become stuck with a standard return of 20 units, equivalent to $p = 0.05$. Perhaps betting on behalf of society requires a consideration of utility which might require the value of 20 to be varied. However, it is difficult to follow the path from decisions to impact on society, so setting a suitable value might be challenging.

REFERENCE

Grünwald, P., de Heide, R. & Koolen, W.M. (2019) Safe testing. Available from: https://arxiv.org/abs/1906.07801

How to cite this article: Smith PA. A distillation of the live chat during the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12671

The following contributions were received in writing after the meeting.

DOI: 10.1111/rssa.12658

Christine P. Chai's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Christine P. Chai
Microsoft Corporation, Redmond, WA, USA

Correspondence: Christine P. Chai, Microsoft Corporation, Redmond WA, USA.
Email: cpchai21@gmail.com

I am delighted to see that testing by betting is an alternative method to *p*-values in terms of communicating statistical findings. Working in the industry and collaborating with non-statisticians, I have received the common advice of 'describe everything in layman's terms, and don't even mention *p*-values in the presentation'. I used to think that the audience has a background in Statistics 101, but the reality is that many do not (or have taken the course but forgotten most of the content). Betting is a more straightforward way to explain how strong the statistical evidence is, because people can imagine that real money is involved. The stronger the evidence against the null hypothesis, the more money people are willing to bet against it.

I am also happy to see discussion on why the *p*-value is not a good measure of statistical evidence, but I do not think that the *p*-value is totally worthless. The four examples in the paper demonstrate that a low *p*-value does not always mean compelling evidence in favour of the alternative hypothesis. But given a sufficiently large number of samples, if we still obtain a high *p*-value, then we can simply conclude that the outcome is insignificant. Therefore, *p*-values can serve as a quick way to remove excessive insignificant variables for dimensionality reduction (Heinze & Dunkler, 2017). The *p*-value cut-off is typically set to a number larger than 0.10, to avoid accidentally removing some important confounding variables. Using *p*-values in variable selection has some success not only in the medical field (Bursac et al., 2008), but also in machine learning (Chai, 2013).

Last but not least, I would like to clarify that I am not trying to defend for the use of *p*-values. Instead, I really appreciate that Dr. Shafer raises concerns about communicating *p*-values and provides testing by betting as an easier method. I also appreciate that Dr. Shafer mentioned the statement on common misunderstandings of *p*-values, which was officially issued by the American Statistical Association (Wasserstein & Lazar, 2016). This article serves as part of the literature that urges people to reconsider the use of *p*-values in hypothesis testing.

REFERENCES

Bursac, Z., Gauss, C.H., Williams, D.K. & Hosmer, D.W. (2008) Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, 3(1), 17.

Chai, C.P. (2013) Facebook account misuse detection—A statistical approach. Master's thesis, National Taiwan University.

Heinze, G. & Dunkler, D. (2017) Five myths about variable selection. *Transplant International*, 30(1), 6–10.
Wasserstein, R.L. & Lazar, N.A. (2016) The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2), 129–133.

How to cite this article: Chai CP. Christine P. Chai's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12658

DOI: 10.1111/rssa.12659

Sander Greenland's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Sander Greenland
Department of Epidemiology and Department of Statistics, University of California, Los Angeles, California, USA

Correspondence: Sander Greenland, Department of Epidemiology and Department of Statistics, University of California, Los Angeles, California, USA.
Email: lesdomes@g.ucla.edu

I promote a perspectivist (toolkit) view in which statistical philosophies are treated as ways of looking at data, and the methods they lead to are then treated as tools whose utility is determined on an application-specific basis. Shafer's betting perspective on testing may be useful for foundational debates, given that disputes over such testing continue to be intense (Wasserstein et al., 2019). To help researchers, however, I believe this perspective will need more elaboration using real examples in which the betting score has a justification and interpretation in terms of study goals, and in which the assumptions in the sampling model are uncertain.

Shafer says 'No one has made a convincing case for any particular choice' of a score derived from a P-value p and then says 'the choice is fundamentally arbitrary'. I think that is quite wrong: some scores have useful interpretations and can be motivated by goals such as information measurement and decision-making. The choice that Shafer displays ($p^{-½} - 1$) appears to have no justification and I find it most unnatural; its absence of history suggests others do as well. Yet there are justifiable choices. One I have seen repeatedly in information statistics and data mining is the surprisal, logworth or S-value $s_b = \log_b(1/p) = -\log_b(p)$, where the log base b determines the scale via $s_b = s_e/\ln(b)$ (Fisher, 1948; Good 1957; Bayarri & Berger, 1999; Boos & Stefanski, 2011; Greenland, 2019; Fraundorf, 2020; Greenland & Rafi, 2020). To explain this choice, I have posted an extended comment on Shafer's article at arXiv (Greenland, 2021).

Regarding terminology, I agree with Shafer insofar as statistics has taken ordinary words like 'significance' and 'confidence' and turned them into overconfident jargon for methods that neglect

crucial information about study problems (Amrhein et al., 2019; Greenland, 2019; Rafi and Greenland, 2020). In the mid-20th century, scientific communities adopted this jargon enthusiastically; yet despite the corruption of the literature it produced, some opinion leaders still defend it vociferously—unsurprising, as to formally account for study problems undermines the significance and confidence that should be assigned to many studies, including their own.

In sum, although Shafer may be offering a valuable new viewpoint, I would need to see how the approach translates into other frameworks and real applications before I could consider using or teaching it, with special attention to what it means when background assumptions are in doubt (Greenland & Rafi, 2021).

REFERENCES

Amrhein, V., Trafimow, D. & Greenland, S. (2019) Inferential statistics are descriptive statistics. *The American Statistician*, 73, 262–270.

Bayarri, M.J. & Berger, J.O. (1999) Quantifying surprise in the data and model verification. In: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. (Eds.) *Bayesian Statistics 6*. Oxford: Oxford University Press, pp. 53–82.

Boos, D.D. & Stefanski, L.A. (2011) P-value precision and reproducibility. *The American Statistician*, 65, 213–221.

Fisher, R.A. (1948) Answer to 'combining independent tests of significance'. *The American Statistician*, 2, 30.

Fraundorf, P. Examples of Surprisal, Available from: http://www.umsl.edu/~fraundorfp/egsurpri.html. Accessed date Aug. 19, 2020.

Good, I.J. (1957, 1983). Some logic and history of hypothesis testing, In: Pitt, J.C., ed. *Philosophical Foundations of Economics*, Dordrecht, Holland, D. Reidel, 149–174. Reprinted as Ch. 14 in Good, I.J. (1983), *Good Thinking*. Minneapolis: U. Minnesota Press, pp. 129–148.

Greenland, S. (2005) Multiple-bias modeling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society, Series A*, 168, 267–308.

Greenland, S. (2019) Valid P-values behave exactly as they should: Some misleading criticisms of P-values and their resolution with S-values. *The American Statistician*, 73, 106–114. https://doi.org/10.1080/00031305.2018.1529625

Greenland, S. (2021). There are natural scores: A comment on Shafer, 'Testing by betting: A strategy for statistical and scientific communication'. Available from: https://arxiv.org/abs/2102.05569.

Greenland, S. & Rafi, Z. (2020). Technical Issues in the Interpretation of S-values and Their Relation to Other Information Measures. Available from: https://arxiv.org/abs/2008.12991.

Greenland, S. & Rafi, Z. (2021). To aid scientific inference, emphasize unconditional descriptions of statistics. Available from: http://arxiv.org/abs/1909.08583.

Rafi, Z. & Greenland, S. (2020) Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Research Methodology*, 20, 244. https://doi.org/10.1186/s12874-020-01105-9, Available from: http://arxiv.org/abs/1909.08579

Wasserstein, R.L., Schirm, A.L. & Lazar, N.A. (2019) Moving to a world beyond "p < 0.05". *The American Statistician*, 73, 1–19. https://doi.org/10.1080/00031305.2019.1583913

How to cite this article: Greenland S. Sander Greenland's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12659

DOI: 10.1111/rssa.12660

Chloe Krakauer and Kenneth Rice's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Chloe Krakauer | **Kenneth Rice**
Department of Biostatistics, University of Washington, Seattle, USA

Correspondence: Kenneth Rice, Department of Biostatistics, University of Washington, Seattle, USA.
Email: kenrice@uw.edu

We agree with Professor Shafer that *p*-values will not disappear, motivating methods which aid their interpretation. Section 3's implied alternative re-interprets *p*-values by feeding them through an arbitrary shrinkage function to give a betting score, which then multiplies the null. As described below we find it simpler but no less productive to define the alternative directly, as a prior. Bayes' 'language of probability' seems the natural one for describing 'hunches', and furthermore brings useful elicitation methods (e.g. O'Hagan et al., 2006) into play.

Our fully Bayesian approach views significance tests as decisions about the sign of a parameter (Rice et al., 2020) for which only alternative, non-null values have support. Our loss functions are not unlike Equation (1)'s bets, although they allow null decisions where nothing is concluded and we incur a fixed penalty, and also 'active' sign decisions—that may be right/wrong, with correspondingly smaller/greater penalties. Our earlier work considers optimal Bayes rule tests, and gives a close analogue of two-sided *p*-values, but the framework can evaluate any test.

Following Rice et al. (2020), we use losses 0/2 for correct/incorrect sign decisions and loss α for null decisions. We evaluate the standard *Z*-test via decision-theoretic *risk* (Robert, 2007, section 2.3). Tests riskier than simply making no decision regardless of the data (with loss and risk α) are termed *futile*, a useful benchmark. Like Shafer, we consider that tests 'merit attention' if they improve on this by fivefold, that is have risk $<\alpha/5$.

For Examples 1–3, where $Y \sim N(\mu, 10^2)$ and α=0.05, we give the standard Z test's risk at `https://kenrice.shinyapps.io/TestingRisk/`. The test is futile with $\leq$ 12.2% power but 'merits attention' with power $\geq$80%—another familiar criterion. We also give prior and posterior risk distributions, propagating uncertainty from Normal priors for μ centred at the various 'hunches'.

Reassuringly, our results agree well with Professor Shafer's results. In Examples 1 and 3, even with the strongest hunches, prior support for the tests meriting attention is not overwhelming. Example 2's large implied target is reflected by >95% prior support for risk $< \alpha/5$, *if* the hunch corresponds to prior SD $\leq$ 5.46.

We can use posterior risk similarly to the betting score. In Examples 1 and 3, underwhelming posterior support for risk $<\alpha/5$ (regardless of hunch strength) indicates that the data do not overcome prior scepticism. In Example 2 reduced posterior support for low risk, versus the prior, points in the other direction to the Z test outcome, as in Professor Shafer's interpretation.

We welcome Professor Shafer's thoughts on quantifying the plausibility of hunches, and how any corresponding calculus differs from that of Bayes.

REFERENCES

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H. & Jenkinson, D.J. et al. (2006) *Uncertain judgements: Eliciting experts' probabilities*. Hoboken, NJ: John Wiley & Sons.

Rice, K., Bonnett, T. & Krakauer, C. (2020) Knowing the signs: A direct and generalizable motivation of two-sided tests. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2), 411–430.

Robert, C. (2007). *The Bayesian choice: From decision-theoretic foundations to computational implementation*. Berlin: Springer Science & Business Media.

How to cite this article: Krakauer C, Rice K. Chloe Krakauer and Kenneth Rice's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12660

DOI: 10.1111/rssa.12661

Kuldeep Kumar's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Kuldeep Kumar
Bond University, Gold Coast, Australia

Correspondence: Kuldeep Kumar, Bond University, Gold Coast, Australia.
Email: kkumar@bond.edu.au

Examples from gambling are quite often used by academic statistician to make the subject more interesting. In fact, the basic tenets of the probability theory were developed by the correspondence between the gamblers. Also, lots of research has been done on various aspects of gambling using statistical tools and techniques. For example, Croucher (2000) used probability intervals to evaluate long-term gambling success. This paper will give a new dimension not only for teaching statistical inference but also effectively communicating the statistical conclusions with betting scores rather than *p*-value approach.

The author has given various interesting protocols, but I am not sure if we can make a protocol for the classification problem. In the classification problems, we quite often take equal cost of misclassification of type I and type II errors but in reality, the cost of misclassification may not be equal. For example, in bankruptcy prediction, the cost of misclassifying a bankrupt company as successful is much more serious as compared to classifying a successful company as bankrupt. In first case, we lose all the money, but in second case, we lose only opportunity to invest and we can take the unequal misclassification cost as 100:1. However, in the case of breast cancer detection, the cost of misclassifying a malignant tumour as benign is much more serious as compared to classifying a benign tumour as malignant and misclassification cost is difficult to evaluate.

REFERENCE

Croucher, J.S. (2000) Using probability intervals to evaluate long-term gambling success. *Teaching Statistics*, 22(2), 42–44.

How to cite this article: Kumar K. Kuldeep Kumar's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12661

DOI: 10.1111/rssa.12662

Tze Leung Lai and Anna Choi's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Tze Leung Lai | Anna Choi
Stanford University, Stanford, CA, USA

Correspondence: Tze Leung Lai, Stanford University, Stanford, CA, USA.
Email: lait@stanford.edu

It is exciting to follow Glenn Shafer's investigations into forecasting, betting, reasoning with uncertainty and foundational issues in probability, beginning with his 1973 PhD thesis at Princeton and culminating in Shafer (1976) on the Dempster–Shafer theory of belief functions, and its evolution during the past five decades to the present paper on betting scores and game-theoretic probability. Betting scores are particularly relevant in this momentous year of intensive global search for COVID-19 vaccines and treatments, and upcoming presidential and congressional elections in the United States, about which pundits keep giving time-varying forecasts of the outcomes while betting markets on presidential election odds have been particularly active, similar to online sports betting markets. Whereas Shafer focuses on the connections between statistical inference and betting, for example using the outcome of a bet against the null hypothesis as 'a simpler way of reporting statistical evidence', Cover and Thomas (1991) have highlighted related connections between betting/gambling and entropy in information theory in their Chapter 6, and complexity (respectively, explanability using probabilities) in their Section 7.2 (respectively, Section 7.10). Shafer mentions in the last sentence of Section 6 the absence in the statistical literature of the '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'. Because the bet/forecast of a current outcome is made on the basis of the past data, there is a natural martingale difference sequence structure which leads to a martingale upon accumulation. Lai, Gross and Shen (2011) have made use of martingale theory to evaluate the statistical performance of probability forecasts and forecasting models/betting strategies, thereby providing such 'cumulative evidential values'.

REFERENCES

Lai, T.L., Gross, S.L. & Shen, D.B. (2011) Evaluating probability forecasts. *The Annals of Statistics*, 39, 2356–2382.
Shafer, G. (1976) *A mathematical theory of evidence*. Princeton: Princeton University Press.

How to cite this article: Lai TL, Choi A. Tze Leung Lai and Anna Choi's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12662

DOI: 10.1111/rssa.12663

Nick Longford's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Nick Longford
School of Public Health, Imperial College London, London, UK

Correspondence: Nick Longford, School of Public Health, Imperial College London, London, UK.
Email: sntlnick@gaia.safeukdns.net

When two courses of action are contemplated, one corresponding to the hypothesis and the other to the alternative, and exactly one of them has to be adopted, betting is the natural and logical response to the uncertainty involved. I part company with the agenda of this paper in its presumption that we have to adhere to the ritual of hypothesis testing. It is widely used; I admit using it because I want to be constructive in my cooperation with colleagues from other specialties, but I regard its every use as a form of compromise with my scientific instincts, principles and perspective. My colleagues' regard for it is no higher, but in their preoccupation with the cycle of FRP (get Funds, do Research and Publish) they are rejecting any new statistical paradigm, just like a housewife or pensioner would not stand up to an aged military dictator. Until the last semblance of an artificial edifice collapses.

Hypothesis testing should be disqualified from statistical practice because it does not have a means of incorporating the consequences of the two kinds of error that may be committed. I grant that it is a great invention and its widespread adoption has contributed to raising the profile of statistics as a unique scientific enterprise, but the steam engine and Karl Marx deserve similar accolades.

The foundation of betting is a currency and the axiom of frugality—that having more of it is universally better. We should use such a currency in our studies and research agendas, be earnest and honest about how much we bet and on what, and what our studies cost and their results are (or are likely to be) worth. Then the arcane calculus of p-values would be replaced by a simple and universally respected additive calculus: losing A and then losing B, in a single instance each, or in expectation, is for all purposes the same as losing A + B. Such a currency, for error, effort, outlay (costs) and knowledge,

may introduce considerable difficulties in establishing (eliciting) their values, akin to constructing the infrastructure for the production and distribution of petrol which we (temporarily) take for granted. But steam powered locomotion is thankfully no longer contemplated.

How to cite this article: Longford N. Nick Longford's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12663

DOI: 10.1111/rssa.12665

Ryan Martin's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Ryan Martin
North Carolina State University, Raleigh, North Carolina, USA

Correspondence: Ryan Martin, North Carolina State University, Raleigh, North Carolina, USA.
Email: rgmarti3@ncsu.edu

Statistics plays a central role in science, so it is imperative that scientists can communicate statistical ideas clearly. I commend Shafer for his efforts to promote his testing-by-betting perspective, but my opinions differ somewhat. Coincidentally, my views have been heavily influenced by Glenn's work in a different direction, including a paper that was read before The Royal Statistical Society almost 40 years ago (Shafer, 1982).

Shafer's betting score proposal is largely motivated by concerns that *p*-values are 'too complicated for effective communication' (Section 1). But the use of *p*-values is statistically well-founded: 'the *p*-value function provides the full statistical story for the particular data value relative to the model' (Fraser, 2013, p. 43). So, if *p*-values are right and 'will never completely disappear' (Section 3), then might a new betting language make scientific communication even more complicated? Would it not be better to understand what *p*-values are and how they should be interpreted?

Fortunately, *p*-values can be understood using Shafer's plausibility theory (e.g. Shafer, 1976). Martin and Liu (2014) showed that, under certain conditions, every *p*-value corresponds to the *plausibility* of the hypothesis, based on the given data, relative to a valid *inferential model* (Martin & Liu, 2013, 2015); similar characterizations of confidence and conformal prediction are given in Martin (2021) and Cella and Martin (2020), respectively. Consequently, the colloquial definition of *p*-values we teach our students, that is, *a p-value measures the plausibility of the hypothesis*, is mathematically

justified. Moreover, interpretation is straightforward: a small p-value indicates the hypothesis is implausible, while a large p-value means the hypothesis is plausible which, for example, does not imply that the data provides evidence supporting the hypothesis.

Concerning efficiency, statistical procedures based on p-values are often best. Since Shafer's betting scores must be smaller than $(p\text{-value})^{-1}$, his betting score-based procedures (Section 4.2) are less efficient than the p-valued-based procedures. Conservative solutions might be warranted under certain circumstances, but should we make across-the-board sacrifices in efficiency for the betting score interpretation?

Finally, I wonder about the use of betting scores when scientists are not actually betting. Shafer's proposed language must lose some of its meaning if those using it are only going through the motions. There are good reasons to make betting a real part of the scientific process (e.g. Crane, 2018) and Shafer's proposal would be much more convincing in such a framework.

REFERENCES

Cella, L. & Martin, R. (2020) Strong validity, consonance, and conformal prediction. Available from: https://researchers.one/articles/20.01.00010.

Crane, H. (2018) The fundamental principle of probability: Resolving the replication crisis with skin in the game. Available from: https://researchers.one/articles/18.08.00013.

Fraser, D.A.S. (2013) Discussion: "Confidence distribution, the frequentist distribution estimator of a parameter: A review" [mr3047496]. *International Statistical Review*, 81(1), 42–48.

Martin, R. (2021) An imprecise-probabilistic characterization of frequentist statistical inference. Available from: https://researchers.one/articles/21.01.00002.

Martin, R. & Liu, C. (2013) Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, 108(501), 301–313.

Martin, R. & Liu, C. (2014) A note on p-values interpreted as plausibilities. *Statistica Sinica*, 24(4), 1703–1716.

Martin, R. & Liu, C. (2015) *Inferential models: Reasoning with uncertainty*, volume 147 of *monographs on statistics and applied probability*. Boca Raton, FL: CRC Press.

Shafer, G. (1976) *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.

Shafer, G. (1982) Belief functions and parametric models (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(3), 322–352.

How to cite this article: Martin R. Ryan Martin's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12665

DOI: 10.1111/rssa.12664

Jorge Mateu's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Jorge Mateu
Department of Mathematics, University Jaume I, Castellón, Spain

Correspondence: Jorge Mateu, Department of Mathematics, University Jaume I, Castellón, Spain.
Email: mateu@uji.es

The author is to be congratulated on a brave, valuable and thought-provoking contribution motivating the use of betting philosophy rather than classical *p*-values. This, for me, means an open door with fresh air when communicating statistical outputs and conclusions to a non-expert audience and to society in general. I liked reading this contribution that stands in its origin but seems to trace a wide and fruitful, while alternative to classical *p*-values, pace towards new times for statistical learning and understanding. However, this is not an easy way, as a change of paradigm is always difficult to setup. This paper is not providing many details on mathematical statistics grounds supporting the betting theory. This is a first battle that this alternative way of thinking has to win: it has to strongly convince the more trained statisticians that betting is as effective and with as many solid grounds as classical *p*-values. I should say that the second battle, an effective use of alternative *p*-values that are better understood, is almost won in a moment when the misuse of *p*-values is largely encountered in numerous influential statistical studies in medicine and the social sciences. I would like to point out key aspects of betting strategies: (a) the betting score blocks any erroneous claim telling us the direction the result points and how strongly; and (b) the uncertainty associated with a large betting score is minimized as much as the certainty provided by a *p*-value is sometimes exaggerated. I would have liked the author to reserve more space in the paper to enhance these aspects, which are somehow hidden and only rise up through some comments.

On a more technical nature, I would like to explore building a betting strategy in those situations where we have a statistical test with an unknown probability distribution under the null or alternative hypothesis. The classical perspective dictates approximating the distribution by simulation. How this could be done under the betting point of view? If we can find a solution to this, there is an open avenue of research when testing complex spatial models that mostly rely on simulations. If betting can provide a simpler and clearer way to explain conclusions in this context to a wider audience, we have advanced the science.

How to cite this article: Mateu J. Jorge Mateu's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12664

DOI: 10.1111/rssa.12666

Stephen Senn's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Stephen Senn
Consultant Statistician, Edinburgh, UK

Correspondence: Stephen Senn, Consultant Statistician, Edinburgh, UK.
Email: stephen@senns.uk

I congratulate Glenn Shafer on a fine and original paper, clearly delivered. I have three comments on connections to the views of RA Fisher.

First, Glenn recommends multiplying betting scores. Fisher recommended multiplicative combination of *p*-values. Section 21.1 of *Statistical Methods for Research Workers* describes how 'to obtain a single test of the aggregate based on the product of the probabilities observed' (Fisher, 1925; p. 99).

Second, Glenn mentions the so-called replication crisis but discusses misuse of *p*-values rather than *p*-values per se as being relevant. This is surely right. Others, in my view, have made the unreasonable leap from showing that a moderately 'significant' value only gives weak evidence against a null hypothesis to assuming that such weak evidence is common but in *Statistical Methods and Scientific Inference* (SMSI) Fisher reminds us 'A man who "rejects" a hypothesis provisionally, as a matter of habitual practice, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is incorrect he will never be mistaken in rejection' (Fisher, 1956; p. 45). Values of *p* are no doubt abused but replacing *p*-values does not, of itself, eliminate abuse.

Third, there is the issue of when one should bet and on what. In the context of gambling the answer is: *bet on whatever you like whenever you like*, *provided you put up the money and register the bet*. What is the standard when we use bets for inference? Fisher was adamant that one had to condition on relevant subsets, writing in SMSI, 'The subject of a statement of probability must not only belong to a measurable set, of which a known fraction fulfils a certain condition, but every subset to which it belongs, and which is characterised by a different fraction, must be unrecognizable' (p. 61). For the practising statistician, conditioning appropriately is an important matter. The problem here is that a perfectly respectable bet (say on the outcome of clinical trial using a *t*-test) might be inferentially irrelevant if further information (such as the value of measured covariates) is available (Senn, 2013). Do any further issues arise by adopting testing by betting? Is there some principle of arbitrage that needs to be observed?

REFERENCES

Fisher, R.A. (1925). Statistical methods for research workers. In: J. H. Bennett (Ed.), *Statistical methods, experimental design and scientific inference*. Oxford: Oxford University.

Fisher, R.A. (1956) Statistical methods and scientific inference. In: J. H. Bennett (Ed.), *Statistical methods, experimental design and scientific inference*. Oxford: Oxford University.
Senn, S.J. (2013) Seven myths of randomisation in clinical trials. *Statistics in Medicine*, 32(9), 1439–1450.

How to cite this article: Senn S. Stephen Senn's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12666

DOI: 10.1111/rssa.12667

Judith ter Schure's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Judith ter Schure
CWI, Amsterdam, The Netherlands

Correspondence: Judith ter Schure, CWI, Amsterdam, The Netherlands.
Email: judith.ter.schure@cwi.nl

Professor Shafer has given a wonderful introduction into the usefulness of 'testing by betting' for interpreting statistical inferences. His talk provided examples of testing weather forecasters and political pundits, showcasing an approach that is more general than conventional statistical tests. Full generality, however, always leaves the applied statistician wondering how to use the approach, which reflects much of the deliberation in the chat.

Applied statisticians, like myself, might find it difficult to specify a bet on an outcome space. On the one hand, we think in types of data (e.g. time-to-event) and accompanying test statistics (e.g. logrank statistic) that are defined in terms of certain model parameters (e.g. hazard ratio). On the other hand, we are used to improving our designs by evaluating the rejection region in a power analysis. Yet instead of thinking of our bets as having power for an *assumed* alternative, we think of them as having power for parameter values specifying effects *we are not willing to miss*.

Bets in terms of these parameters provide a clear instruction to develop new statistical methodology (e.g. the *t*-test, contingency table test and logrank test, which we were able to design (Grünwald et al., 2019; Grünwald et al., 2020): bets should make a profit for parameters of minimal interest, but not for any distribution in the null hypothesis. Nevertheless, staying close to statistical practice does call for a slight relaxation of 'opportunistic betting'. Unlike the Bayesian, we might not wish to use prior knowledge, or optimize our 'prequential' forecasts. Instead of focusing on the parameter values

we assume most likely to be true—based on prior knowledge—we might wish to bet on the values for which, if they are true, we most dearly want this to show in our betting profit.

Our bets might deliberately miss out on profit for effects smaller than our smallest effect size of interest. Hence the choice of bets poses a challenge for retrospective meta-analysis (which we try to address in the ALL-IN meta-analysis philosophy (ALL-IN meta-analysis): each study can have its own protocol and effect of minimal interest.

I am already convinced that it will serve statistics well to replace p-values by bets, and power analyses by implied targets. But how do we know whether practitioners actually find this more intuitive? As an applied statistician, I will keep that question from the chat in mind as well, and see if I can design an experiment to test it.

REFERENCES

Grünwald, P.D., de Heide, R. & Koolen, W. (2019) Safe testing. *arXiv preprint arXiv:1906.07801*.
Grünwald, P.D., Ly, A., Pérez-Ortiz, M.F. & ter Schure, J. (2020) Safe logrank test: arXiv preprint arXiv:2011.06931.
ALL-IN meta-analysis: soon to appear as a preprint.

How to cite this article: Schure J. Judith ter Schure's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12667

DOI: 10.1111/rssa.12668

Paul Vos's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Paul Vos
East Carolina University, Greenville, NC, USA

Correspondence: Paul Vos, East Carolina University, Greenville, NC, USA.
Email: vosp@ecu.edu

I appreciate Professor Shafer's effort to communicate statistical inference to a wide audience and I agree that simplicity is key. I disagree that the *p*-value is too complicated. It is the standard presentation of probability involving infinite sequences of hypothetical repeated samples that is too complicated. The *p*-value is a probability that can be described simply as a proportion. Classical probability affords the simplicity required for effective communication.

I do not think about inference initially in terms of 'certain phenomenon *Y*'. Much of statistical inference begins with a finite population. The population refers both to individuals and numeric quantities obtained by measuring individuals. This can be conceptualized in terms of a bowl containing a ball for each individual with the measured value(s) written on the ball.

For every population, there is the n-sampling distribution which consists of every subset of size n. The elements of this distribution are ordered by a real-valued function T whose definition depends on the inference question. The distribution of T is the multiset $\lfloor T \rceil$ that can be thought of as a bowl with one ball for each element of the n-sampling distribution. Numerically, the p-value is the proportion of balls in $\lfloor T \rceil$ that are as or more extreme than the ball corresponding to the observed sample. For this proportion to be a probability, only one randomization is required—the one used for the actual data.

The p-value described above belongs to probability theory. For inference, a model M_θ for the population is hypothesized and the resulting multiset denoted $\lfloor T \rceil_\theta$. The support of M_θ may be infinite but can be discretized so that $\lfloor T \rceil_\theta$ is finite. This multiset need not be constructed; the point here is that $\lfloor T \rceil_\theta$ closely approximates M_θ so that classical probabilities can be used to communicate to a wide audience. As θ varies there are infinitely many multisets and so infinitely many proportions none of which require randomization or sampling. All of these proportions become probabilities by virtue of a single randomization. Additional details appear in Vos and Holbert (2019).

Penrose (2005, pp. xvi–xviii) describes how the mathematical description of a fraction (as an equivalence class of ordered pairs of whole numbers) can obscure the intuitive notion associated with this concept. I fear we have done the same for probability. Insisting on hypothetical repetitions has obscured the simple notion of probability as a proportion.

REFERENCES

Penrose, R. (2005) *The road to reality: A complete guide to the laws of the universe*. New York: A.A. Knopf. ISBN: 0679454438.

Vos, P. & Holbert, D. (2019) Frequentist inference without repeated sampling. *arxiv: 1906.08360 [stat.OT]*

How to cite this article: Vos P. Paul Vos's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12668

DOI: 10.1111/rssa.12669

Ruodu Wang's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Ruodu Wang
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada

Correspondence: Ruodu Wang, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada.
Email: wang@uwaterloo.ca

In this comment, I will refer to Shafer's betting scores as *e-values* (following Vovk & Wang, 2020b). I will use *e*-values as the abstract term in a measure-theoretic setting encompassing specific interpretations including betting scores, likelihood ratios and (stopped) test supermartingales; see Table 1 for an analogy of *e*-values and *p*-values.

e-values may be superior in (but not limited to) situations where (i) precise *p*-values are not easy to compute (e.g. unverifiable asymptotics); (ii) experiment results need to be sequentially updated or monitored; (iii) communication of likelihood ratios or betting scores is important; or (iv) model ambiguity, such as complicated dependence, exists in the data. These issues are (partially) discussed by Shafer; see also Shafer et al. (2011), Grünwald et al. (2020) and our working papers on *e*-values posted on alrw.net/e.

e-values are natural to combine, making them convenient in both the accumulation of evidence (single hypothesis) and the selection of findings (multiple hypotheses). The most basic way of combining *e*-values is averaging: The arithmetic mean of *e*-values is an *e*-value (an *e-merging function*) without any assumptions on the dependence structure of the input *e*-values, and it essentially dominates (Vovk & Wang, 2020b) all other symmetric *e*-merging functions. Quite the contrary, the structure of symmetric *p-merging functions* (combining *p*-values) is much more complicated; see, for example, Vovk and Wang (2020a). Even if we are not really interested in *e*-values and only interested in combining *p*-values, *e*-values still appear as a powerful technical tool: All admissible ways of merging *p*-values need to go through a calibration to *e*-values via a duality theorem in optimal transport (Vovk et al., 2020).

Methods for combining *p*-values are widely used in testing multiple hypotheses. In the context of false discovery control, methods based on *e*-values lead to encouraging first results, similar to the *p*-value-based procedures of Benjamini and Hochberg (1995) and Goeman and Solari (2011); see Vovk and Wang (2019) and Wang and Ramdas (2020).

Combined *e*-values can often be improved when the input *e*-values are independent or, more generally, *sequential* (a sequence where each element is a valid *e*-value conditional on previous ones). This is the situation that Shafer considers in his betting games; test martingales of Shafer et al. (2011) appear as the product processes of such *e*-values. Other useful merging functions

TABLE 1 An analogy of p-values and e-values. In the table, $\mathbb{P}$ is a null probability measure, $\mathbf{X}$ is the vector of data, $T(\mathbf{X})$ is any test statistic, T' is an independent copy of $T(\mathbf{X})$, $\mathbb{Q}$ is any probability measure, M is a test supermartingale and τ is a stopping time

	Requirement	Specific interpretation	Representative forms	Keyword
p-value P	$\mathbb{P}(P \leq \alpha) \leq \alpha$ for $\alpha \in (0,1)$	Probability of a more extreme observation	$\mathbb{P}(T' \geq T(\mathbf{X})\,|\,\mathbf{X})$	(conditional) probability
e-value E	$\mathbb{E}^{\mathbb{P}}[E] \leq 1$ and $E \geq 0$	Likelihood ratios, stopped martingales and betting scores	$\mathbb{E}^{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\middle|\mathbf{X}\right]$ $\mathbb{E}^{\mathbb{P}}[M_\tau\,|\,\mathbf{X}]$	(conditional) expectation

for sequential e-values include mixtures of U-statistics and those obtained from stopping other supermartingales.

ACKNOWLEDGEMENTS

I am grateful to Vladimir Vovk for advice and insightful discussions.

REFERENCES

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289–300.

Goeman, J.J. & Solari, A. (2011) Multiple testing for exploratory research. *Statistical Science*, 26, 584–597. Correction: 28:464.

Grünwald, P., de Heide, R. & Koolen, W.M. (2020) Safe testing. *arXiv:1906.07801 [math.ST]*, arXiv.org e-Print archive.

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

Vovk, V., Wang, B. & Wang, R. (2020) Admissible ways of merging p-values under arbitrary dependence. *arXiv:2007.14208 [math.ST]*, arXiv.org e-Print archive.

Vovk, V. & Wang, R. (2019) True and false discoveries with e-values. *arXiv:1912.13292 [math.ST]*, arXiv.org e-Print archive.

Vovk, V. & Wang, R. (2020). Combining p-values via averaging. *Biometrika*, 107(4), 791–808.

Vovk, V. & Wang, R. (2021) E-values: Calibration, combination, and applications. *Annals of Statistics*. To appear. *arXiv:1912.06116 [math.ST]*, arXiv.org e-Print archive.

Wang, R. & Ramdas, A. (2020) False discovery rate control with e-values. *arXiv:2009.02824 [math.ST]*, arXiv.org e-Print archive.

How to cite this article: Wang R. Ruodu Wang's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12669

DOI: 10.1111/rssa.12670

Priyantha Wijayatunga's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Priyantha Wijayatunga
Department of Statistics, Umeå University, Umeå, Sweden

Correspondence: Priyantha Wijayatunga, Department of Statistics, Umeå University, Umeå, Sweden.
Email: priyantha.wijayatunga@umu.se

Here I make two points on hypothesis testing. The p-value is the conditional probability of seeing the test statistic having realizations the same or more extreme from what already available data have shown, given the null hypothesis H_0. Since the word 'conditional' is often omitted in text books, many practitioners believe that the p-value is just an unconditional probability, thus tend to misuse it. One may argue that if we obtain the correct p-value often. Suppose we test if the mean of a normal population, say, μ is positive with a sample of observations with size n from the population. For unknown population variance, the test statistic T calculated from the sample has a t-distribution with degrees of freedom $(n-1)$ under H_0. Therefore, the p-value is $p = P\{T \geq t_o \,|\, H_0 \text{ is true}\}$ where t_o the observed test statistic value. One can make a reasonable argument that the p-value that we calculate is often smaller than its true value for the application since, for example, we assume that our data are a random sample; our observed p-value,

$$\begin{aligned}\hat{p} &= P\{T \geq t_o \text{ and data are a random sample} | H_0 \text{ is true}\} \\ &= P\{\text{data are a random sample}\} P\{T \geq t_o | H_0 \text{ is true}\} \leq p.\end{aligned}$$

So, in practice, we may be rejecting the null hypothesis more often than what it should have been. This means that we should inflate our calculated p-value to a certain degree.

Second, consider a test (see Sprenger, 2013); out of 104,490,000 Bernoulli trials, 52,263,471 are successes and 52,226,529 are failures, therefore observed probability of success is 0.5001768. For testing if the true value of it is 0.5, we get a p-value that is lower than 0.01. Therefore, it is rejected at 0.01. The standard error of the estimate of the probability of success is 0.00004891394 that is almost equal to its value under null hypothesis. For the purpose of deciding if the true probability of success is 0.5, do we need to do a hypothesis test, since the empirical estimate is almost the same as the test value, and the standard error of the estimate is practically zero? What is the purpose of doing a test under these circumstances? If we take that the standard error to be zero, then we should accept

that the value of the estimate is 0.5. We do not need hypothesis tests to communicate the statistical result in this case. The hypothesis tests are only mathematically objective procedures that have no subjective opinions embedded in them. However, use of any statistical result is often subjective or contextual!

REFERENCE

Sprenger, J. (2013) Testing a precise null hypothesis: The case of Lindley's paradox. *Philosophy of Science*, 80(5), 733–744.

How to cite this article: Wijayatunga P. Priyantha Wijayatunga's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12670

DOI: 10.1111/rssa.12672

Author's reply to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer

Glenn Shafer
Rutgers University, Newark, New Jersey, USA

Correspondence: Glenn Shafer, Rutgers University, Newark, New Jersey, USA.
Email: gshafer@business.rutgers.edu

The author replied later in writing, as follows:

I am gratified by this bounty of thought about my paper and my talk. I am also grateful to the organizers of the meeting, who worked so hard to make it work in this time of remote communication, and to Philip Dawid and Frank Coolen for initiating the vote of thanks.

I offer special thanks to the discussants who had already contributed directly to the paper's ideas: Dawid, Vladimir (Volodya) Vovk and Peter Grünwald. Phil's work on probability forecasting in the 1980s and 1990s, especially his prequential principle, was seminal, and he has been a constant presence in the further development of game-theoretic probability by Volodya and myself. Peter's work is a more recent influence, but the writing of the paper was spurred by his and my efforts to understand each other, and his comments on the paper mark the success of those efforts.

Thanks also to all the other discussants: old friends, new friends and others I hope to meet. My greatest delight is seeing a new generation of researchers who have been putting the ideas of my paper, sometimes differently named and motivated, to work.

Testing pundits and weather forecasters. As Tze Leung Lai and Anna Choi point out, betting scores are particularly relevant today, when so much attention is paid to time-varying probability forecasts for weather, elections and sports. Our weather forecasts change hourly. My paper assumes that only one

probability forecast is made for each outcome. But I began my talk by pointing out that we can also test by betting when the forecaster changes their forecast repeatedly before an outcome is realized or revealed.

In January 2020, the widely followed statistician and pundit Nate Silver announced that he would post and regularly update probabilities for who would become the Democratic nominee for the US presidency. How could we have tested Silver's successive probability distributions? By betting. On Day 1, using play money because Silver had not offered to bet, we could have bought a nonnegative payoff to which Silver's Day 1 distribution assigned expected value 1. On each following day we could have 'rebalanced our portfolio', selling back to Silver at his new prices the payoff we had bought on the previous day and buying a new one with the proceeds. Had we multiplied our initial 1-unit stake by a large factor when the contest was over, we could claim to have discredited Silver's forecasting.

Shafer and Vovk (2001) suggested using this method to test the efficiency of financial markets. How else can you do it? The academic community in finance has not found any alternative. They once tried to address the issue by testing models that account for the observed prices. But this led nowhere; it merely created what financial economists have called 'the joint-hypothesis problem'. Eugene Fama (1991) hoisted a white flag with these words:

> Ambiguity about information and trading costs is not, however, the main obstacle to inferences about market efficiency. The joint-hypothesis problem is more serious. Thus, market efficiency per se is not testable. It must be tested jointly with some model of equilibrium, an asset-pricing model.

Twenty years ago, Leo Breiman warned us that the culture of prediction was overtaking the culture of stochastic data models (Breiman, 2001). We now see, in business, government and our daily life, a flood of numerical predictions, many produced by algorithms (e.g. neural networks and physical models) that bear only a tortured resemblance to stochastic models. Testing by betting provides common ground, a starting point for teaching and understanding that connects significance tests, estimation and likelihood with other algorithms of data science.

Perhaps this broader perspective can be advanced by replacing the terms 'hypothesis testing' and 'significance testing' with 'prediction testing'. Even Nick Longford, who wants us to abandon the 'ritual of hypothesis testing' in favour of a calculus that sums up expectations of costs and benefits, may sometimes want to check how well his expectations are predicting what happens.

Choosing among bets. Philip Dawid asks what we should do when we have no particular alternative Q in mind but think of several bets we might like to make. Can we make them all, find the resulting maximum, and make some sort of multiplicity adjustment? Yes. Any nonnegative increasing function of the maximum that has expected value 1 will provide the needed adjustment. But an easier and usually better way of combining bets is to average them; any weighted average of a collection of bets is itself a bet. As Ruodu Wang points out, this requires no assumptions about the dependence structure and essentially dominates any other method of combination.

Averaging bets two bets against P, say S_1 and S_2, is equivalent to averaging the two implied alternatives S_1P and S_2P. The idea of averaging possible alternative distributions is very familiar. When we when average alternative distributions with weights proportional to prior probabilities, the resulting betting score is identical with what is often called the *Bayes factor* for testing the simple null hypothesis P. (See, e.g. Kass & Raftery, 1995, p. 776. The name is more often used for the inverse of this quantity, which is Harold Jeffreys's K.) We should not, however, equate betting scores with Bayes factors. Some betting scores are not Bayes factors. These include any betting score obtained by multiplying betting scores against successive predictions when the theory, forecaster or pundit making the predictions did not give a

joint distribution in advance and hence the bettor did not make a joint bet in advance. As Vladimir Vovk notes, some Bayes factors (namely, many for testing composite null hypotheses) are not betting scores.

Aaditya Ramdas, Vovk and Wang report numerous important examples of the effectiveness of averaging bets. Ramdas notes that he has shown that such averaging can lead to inferences that are, in some practical sense, universal. The idea of averaging all bets that we can compute in order to obtain a universal bet goes back to Jean Ville, who noted that the universal bet thus obtained is not itself computable (Bienvenu et al., 2009). For Ville, this impossibility of a bet that it is both truly universal and implementable was a reason to use axioms rather than game theory as the mathematical foundation of probability. But from a more practical point of view, it may mean only that what is universal for one purpose may not be universal for another.

Judith ter Schure provides important insights about the choice of bets from the perspective of an applied statistician doing meta-analysis. If I understand correctly, she proposes to consider not an alternative to P that she considers most possibly true but rather one that departs from P in ways most damaging if P is used as an input for decisions. This can be achieved most readily by choosing a test statistic T that measures the unwelcome departures, perhaps a statistic already widely used or a simple modification of such a statistic. If T is nonnegative and bounded, then $T/\mathbf{E}_P(T)$ may serve as an appropriate bet. Averaging bets may also have a role to play here. Suppose, for example, that $Y = (W,Z)$. Suppose the statistician is most concerned about P's predictions of Z and worries that that these predictions may err in certain ways depending on W. In this case, she might begin by choosing bets S_w against P's conditional distribution for Z given w. Then she can obtain a bet S by averaging the S_w using P's marginal for W. It will be interesting to see these ideas put into practice.

Betting scores and e-values. My paper's notions of bet and betting score are very old, and a variety of names have been given to these objects. In 1993, Vladimir Vovk and Vladimir V'yugin wrote $p(\omega)$ instead of $S(y)$ and called p an *impossibility measure* (Vovk & V'yugin, 1993, p. 257). Testing by betting is the central idea of Shafer and Vovk (2001, 2019), but these books did not settle on a single name and symbol for the factor by which Sceptic's capital is multiplied. After the 2019 book was completed, Vovk and I introduced new names for this factor. I introduced *bet* and *betting score* in an early version of the paper under discussion (Shafer, 2019a). Vovk (2019) introduced *e-value* for what I am calling a betting score, and Vovk and Ruodu Wang subsequently settled on *e-variable* for what I am calling a bet (Vovk & Wang, 2019).

The name *e-value* has proven attractive, because it evokes only the standard notion of expected value, with no explicit reference to betting and no hint of any game-theoretic excursion outside familiar territory. In the discussion, we see *e-value* used by Peter Grünwald and Aaditya Ramdas, as well as by Vovk and Wang. The name *e-variable* has been used less. Vovk and Grünwald both distinguish systematically between *e-variable* and *e-value*. But Wang uses *e-value* for both concepts, and Ramdas first distinguishes *e-value* from *e-variable* but then calls a martingale an e-value. These variations are not surprising; variable/value distinctions generally tend to be unstable in mathematical conversation. I am hoping that the distinction *bet/betting score*, being less abstract, will be easier to bring to mind in settings where it is helpful. A bet is an action; the betting score is the result of the action.

In one respect, the notion of a *betting score* is more general than Vovk's notion of an *e-value*. In my paper, I emphasized the case where one scientist makes a bet S_1 on the outcome of her experiment and then, perhaps because $S_1(y_1)$ is large but only moderately so, the same or a different scientist does a newly devised experiment to test the same hypothesis, making the bet S_2. This, I proposed, yields a betting score $S_1(y_1)S_2(y_2)$, even though no corresponding bet or strategy for betting had been declared at the outset. No one had a joint distribution for Y_1 and Y_2 at the outset, because no one had a model for whether and how some someone might react to the first result and devise and perform the new experiment that created Y_2. So we have a betting score with no single bet, whereas an e-value must always be the value of an e-variable.

Some readers may prefer to eliminate this divergence by positing that some demigod, who somehow had a joint distribution for what experiments would be performed but did not know how they would come out, made a bet S whose payoff turned out to be $S_1(y_1)S_2(y_2)$. I am not tempted by this fantasy, and I question its usefulness in scientific communication. I propose that we instead rely on the simple principle that a forecaster or theory is discredited when and to the extent that a person or community multiplies an initial stake by a large factor by betting against its predictions, no matter whether there was only one bet or many in succession, and no matter how each successive bet was chosen.

When S is a bet at prices given by P, Markov's inequality tells us that

$$P\left(S \geq \frac{1}{\alpha}\right) \leq \alpha. \tag{1}$$

The intuition that $S(y)$ should not be large if P is correct is then supported by the principle that an event of small probability should not happen. This principle, sometimes called *Cournot's principle* (Shafer & Vovk, 2006), underlies all 'frequentist' statistical methods. But when spoken aloud, it evokes debate from students and philosophers. Is not what happens always an event of small probability? A good answer is that we consider only simple events that are declared in advance. If our students remain unsatisfied, we hasten to change the subject with words like 'significance' and 'confidence'.

Implicit in my paper is the proposal that we turn our thinking about (1) upside down, taking as basic not the contested principle that an event of small probability should not happen but the principle that multiplying my money by a large factor discredits P's probabilistic predictions. This has multiple advantages as a strategy for communication. The public can understand it. It makes clear the need for a declaration in advance, for I cannot make a bet without declaring it. It puts just the right burden on the argument that failure to multiply my capital would be evidence for the hypothesis or theory being tested. (Am I so knowledgeable and effective, and the data so extensive, that no one will be able to refute the hypothesis if I cannot?) It requires no discussion of what probability really means and no fantasies when we think about $S_1(y_1)S_2(y_2)$ as evidence against a theory even though it is not the realized value of a single bet.

When there is not a single bet, we are outside the framework of a single probability distribution. In this sense, we are outside measure-theoretic probability. As Vovk explains, the natural habitat for 'betting scores' is game theoretic, the natural habitat for 'e-values' measure theoretic. Shafer and Vovk (2019) show that game-theoretic probability has the same rigor as measure-theoretic probability and that the two are equivalent in essential respects. But working with betting scores does not require a study of game-theoretic probability, any more than working with significance tests requires a study of measure-theoretic probability.

Stephen Senn notes the analogy between multiplying betting scores and R. A. Fisher's multiplication of p-values. The analogy is only partial, because the product p_1p_2 of two p-values p_1 and p_2 is not itself a p-values. You must treat p_1p_2 (or $-\ln p_1p_2$, as Fisher preferred) as a test statistic and calculate a new p-value from it. But the broader logic is parallel. Fisher did not give a fig about measure theory, and he did not worry about the absence of a comprehensive probability distribution, posited in advance, for the existence and outcomes of multiple experiments.

Confidence and warranty. Philip Dawid asks, 'does a $1/\alpha$-warranty set have any property analogous to the coverage property of a classical confidence interval?' Yes, at least in Protocol 5, the context in which I defined $1/\alpha$-warranty in my paper. In this protocol, where each parameter value defines a joint probability distribution for all the observations, a $1/\alpha$-warranty set is the same thing as a $(1-\alpha)$-confidence set. Only the name and the interpretation are different. I should have said this in the paper.

Why is $1/\alpha$-warranty the same as $(1-\alpha)$ confidence in Protocol 5? Because once Sceptic knows θ, the protocol reduces to Protocol 2, and there Sceptic has a strategy that multiplies his money by $1/\alpha$ for all $(y_1, \ldots, y_N) \in W$ if and only if $\alpha \geq P(W)$. Paraphrasing:

> $P(W)$ is the least number α such that Sceptic can multiply his money by $1/\alpha$ whenever W happens.

This is the game-theoretic definition of probability (Shafer & Vovk, 2019, sections 2.1,7.3). The more general statement for expected value was Christiaan Huygens's definition of the value of an expectation (Shafer, 2019c), but today's textbooks no longer treat it as a basic principle. It can be derived from today's principles using Markov's inequality and the reasoning in my paper in the paragraph after Protocol 2. To extend the reasoning to protocols where N is random, use Ville's inequality rather than Markov's (Shafer & Vovk, 2019, p. 49).

Being identical to confidence intervals, warranty sets can share all their problems, including the problem of relevant subsets mentioned by Stephen Senn. But insights provided by the picture of sequential betting may sometimes help us choose warranty strategies (= confidence procedures) that mitigate some of these problems. These insights can be used even when the data are obtained as a batch rather than sequentially (Waudby-Smith & Ramdas, 2020).

The identity between $(1-\alpha)$-confidence and $1/\alpha$-warranty holds in any context where confidence sets can be defined. But the notion of $1/\alpha$-warranty also extends to the situation where a sequence of experiments is not initially planned or even contemplated. There we can obtain a final $1/\alpha$-warranty set by supposing that Sceptic, who knows the value of θ from the outset of the first experiment, uses his capital at the end of each experiment to begin betting in the next. The statistician's final $1/\alpha$-warranty set will then be the set of θ for which the final capital after all the experiments is less than $1/\alpha$. Equating the warranty set with a confidence set in this situation would again require a very special demigod.

The coverage property that Dawid mentions requires that a confidence set cover the true parameter value with probability at least $1-\alpha$. Some statisticians, including Antoine Augustin Cournot (Cournot, 1843, section 108) but not including Dawid, have further explained this by saying that if someone (another demigod!) were to repeat the experiment and the calculation many times in exactly similar circumstances, the true parameter value would be in the calculated set at least $1-\alpha$ of the time. Paul Vos pins our difficulties in communicating about probability on this talk about hypothetical repetition. I agree that such talk is no longer helpful in the ways it might have been in 1843. It misleadingly privileges identical repetition, directing our attention away from the case of a diverse unplanned sequence of tests.

All warranty sets do have a game-theoretic coverage property: a $1/\alpha$-warranty set will fail to cover the true parameter value with *game-theoretic probability* at most α. However, this is not a deep statement. By the definition of game-theoretic probability, an event has game-theoretic (upper) probability α or less when there is a betting strategy that multiplies the capital risked by $1/\alpha$ or more when the event happens.

Paul Smith reports that a familiar proposal popped up in the live chat: we should stop testing subsets of parameter spaces in favour of confidence intervals. One shortcoming of this proposal is that it applies only to models that specify a class of probability distributions (parametric or nonparametric), ignoring all other situations where we want to test predictions. Another difficulty, and the reason the proposal has never taken root, is that confidence intervals inevitably give birth to p-values and their abuses. Confidence intervals were first widely used after Joseph Fourier explained how to calculate them for the difference between two proportions, using error probabilities 1/2, 1/20, 1/200, etc. But who will refrain from asking which of the intervals do and do not contain zero? Jules Gavarret argued for fixed-level significance testing for the difference between the success rates of two medical

treatments. But p-values were of course more popular. By 1843, Cournot was denouncing the resulting abuses. See Shafer (2019b) for details.

Betting strategies for parametric statistics. There is a large literature in parametric statistics concerning how tests of different parameter values might cohere. How should the way we test θ_1 be related to the way we test θ_2? Should we seek or expect to find a test of θ_1 that is best both when θ_2 is true and when θ_3 is true? Philip Dawid asks whether there are general principles for answering these questions when we test by betting.

Dawid uses my Protocol 7 as an example. It was structured so as to suggest a connection between Sceptic's betting strategies for the different values of μ: the bet on the error e_n should not depend on μ, which Sceptic knows, but only on the preceding errors $e_1,\ldots,e_{n-1}$. Is this a matter of principle? I think not. I exploited the group structure to get a simple betting strategy for Sceptic, but whether the statistician chooses this strategy should depend, I think, on the statistician's hunches about how the model might err. If the statistician has different hunches for different μ, or perhaps hunches that depend on other information (signals as in my Protocol 4), then she should take this into account rather than respect the group structure.

Similarly, we will often fail to find a test of θ_1 that is best against all alternatives. In protocols such as Protocol 5, where each parameter value specifies a complete probability distribution, the for each parameter value, the optimal bets $P_{\theta_2}/P_{\theta_1}$ and $P_{\theta_3}/P_{\theta_1}$ will generally be different.

Xiao-Li Meng's and Peter Grünwald's comments clarify the relationship between betting scores and Bayes factors for composite null hypotheses. In the case of a composite null hypothesis Θ_0 and a simple or composite alternative Θ_1, a Bayes factor is a random variable S of the form $S(Y)=Q(Y)/P(Y)$, where P is the weighted average of the distributions in Θ_0 obtained using a prior distribution μ_0 on Θ_0, and Q is similarly obtained using a prior distribution μ_1 on Θ_1. We have $\mathbf{E}_P(S) = 1$, and so S is a bet for testing P. But we did not ask for a bet that tests P. We wanted to simultaneously test all the P_θ with $\theta \in \Theta_0$. The relation $\mathbf{E}_P(S) = 1$ does not imply that S is such a bet.

Grünwald reports that he and his colleagues have shown that for every distribution μ_1 over the alternative, we can choose a distribution μ_0 over the null such that the resulting Bayes factor does test all $\theta \in \Theta_0$. This is nice generalization of what we know about a simple null hypothesis P: for every simple alternative Q there is a bet (namely Q/P) that tests P and whose value will be a Bayes factor for testing P.

Meng reports that his research with colleagues reveals a different connection between betting scores and Bayes factors for composite null hypotheses. Suppose a parameter θ, for which we have a prior π, indexes both the null and the alternative: the null is a class $(P_\theta)_{\theta\in\Theta}$; the alternative is a class $(Q_\theta)_{\theta\in\Theta}$. Write P and Q for Y's marginal under the null and alternative respectively. Then the Bayes factor $Q(y)/P(y)$ is the expected value under π of the betting score $Q_\theta(y)/P_\theta(y)$ that we could have calculated had we known θ.

In some problems, including Harold Jeffreys's problem of testing whether additional parameters are needed, it seems reasonably natural to index the null and alternative with a common parameter. In others, such an indexing may be rather contrived. But in any case, a subjective expected value for an unknown betting score against the predictions of a partially known hypothesis is not necessarily a betting score against them, and its value as a test result may be questioned. If we use my device of having the statistician stand outside a betting protocol in which Sceptic does the betting, we may be unimpressed when the statistician announces a high subjective expected value for the betting score.

Betting scores and p-values. A fixed significance level α and a numerically equal p-value p have different meanings and carry different weight. Rejection at level 0.01 means that an event selected in advance and alleged to have this small probability has happened. In the case of a p-value 0.01, the

event alleged to have this probability was not selected in advance; only a class of events, the tail events for a particular test statistic, was selected in advance. So the p-value carries less weight. The two numbers are on different scales. It is reasonable to ask for a convention for shrinking p to fit it onto the α scale, while acknowledging that such a convention must be largely arbitrary. Because $1/\alpha$ is the payoff when the fixed-level test is interpreted as a bet, the scale for betting scores is the same as that for $1/\alpha$. So the task is to choose a convention for translating p to the $1/\alpha$ scale.

I am delighted to learn, from Vladimir Vovk, that Harold Jeffreys's rule of thumb for comparing p-values with Bayes factors agrees closely with my suggestion, $p \mapsto p^{-1/2}-1$, for the key values $p = 0.05$ and $p = 0.01$. It is also notable that in the same appendix where Jeffreys offers his rule of thumb, he suggests that a Bayes factor of 10^{-2} can be considered decisive. Inverting this, we obtain a betting score of 100, corresponding under the mapping $p \mapsto p^{-1/2}-1$ to p=1/10,201. This number would not surprise Joseph Fourier, who treated a large-sample confidence interval as conclusive when the error probability had this order of magnitude.

The main purpose of the mapping $p \mapsto p^{-1/2}-1$ is to facilitate conversation between people using different methods of inference; it may help one person understand roughly the meaning of numbers reported by another. But if you want a betting score, there is no good reason to calculate a p-value and then map the p-value to a betting score. It may be better to work directly with the test statistic T from which a p-value might be calculated. Usually T is chosen so that it tends to be larger than P expects when P is wrong in a way that concerns us. Even when we find it difficult to formulate a particular alternative Q for Y, we may be able to specify a range of values of $T(Y)$ that we are most concerned about, and this may help us fashion a bet $S(T(Y))$.

Sander Greenland writes, quite reasonably, that other mappings from [0,1] to [0,∞] are appropriate when other study goals are pursued. When measuring information, for example probabilities are often transformed using $p \mapsto -\log_2 p$. The study goal of my paper is testing and combining tests, not information measurement. As I wrote in my section 3, I need a function f such that $f(p)$ has expected value 1 when p is uniformly distributed between 0 and 1. The function $p \mapsto -\log_2 p$ does not satisfy this condition. The function $p \mapsto -\ln p$ does, but $p \mapsto p^{-1/2}-1$ and other more complicated functions in the literature referenced in (Shafer & Vovk, 2019, section 11.5) come closer to established opinion about the comparison of p-values with likelihood ratios and Bayes factors. Here are a few comparisons, using the inverse functions $S \mapsto \exp(-S)$ and $S \mapsto 1/(S+1)^2$.

S	$S \mapsto \exp(-S)$	$S \mapsto 1/(S+1)^2$	Jeffreys's rule of thumb
$10^{1/2}$	0.04	0.06	0.05
10	5×10^{-5}	0.008	0.01
15	3×10^{-7}	0.004	
$10^{3/2}$	2×10^{-14}	0.0009	

Harold Jeffreys called values of his K between 10^{-1} and $10^{-3/2}$ 'strong' evidence (Jeffreys, 1961, p. 432). R. A. Fisher never made an equally relevant comparison, but he did state that parameter values with likelihood less than 1/15 of the maximum likelihood are 'open to grave suspicion' and are 'definitely unlikely' (Fisher, 1956, sections III.6 and V.7).

There is an interesting connection between the problem of shrinking a p-value and the problem of adjustment when Sceptic is supposed to announce his final capital after a sequence of bets but instead announces the maximum he attained. As it turns out, the mappings that seem acceptable for the two cases are the same. This was demonstrated in Dawid et al. (2011), and it provides a partial answer to a question Dawid himself asks now: can betting scores sometimes be adjusted to account for some information not being reported?

Simplicity. Participants in the live chat suggested that experiments might help us decide whether testing by betting is simpler than conventional testing, and Judith ter Schure promised to think about how to design relevant experiments.

Perhaps the most crucial choice will be the selection of participants. Will they be highly trained statisticians? Poker players, as Philip Dawid playfully suggests? Scientists who use statistical tests? Students in statistics classes? Or perhaps teenagers?

I did not play poker as a teenager, but I remember that my classmates, when disputing each others' predictions, readily used betting taunts: 'Wanna bet?', 'Put your money where your mouth is'. Imagine making these reports to teenagers:

- Prof Shafer tested your app's rainfall predictions over the past year by betting against them and turned \$1 into \$10. He concluded that the app is not doing a good job.
- Prof Shafer constructed a statistical model for your app's rainfall predictions over the past year. The app's predictions were inaccurate by an amount he would have expected only 1% of the time. He concluded that if his model is right, the app is not doing a good job.

Which report would the teenager be more likely to remember? Which would she be able to repeat to a classmate? If you teach statistics, which do you think your students would be able to repeat accurately?

Arthur Paul Pedersen asks what would be left of my paper's contribution if my claims about simplicity were jettisoned. Other participants, especially Peter Grünwald, Aaditya Ramdas, Vladimir Vovk and Ruodu Wang, answer Pedersen's question by pointing to numerous applications where simplicity is not the only salient advantage of testing by betting. But simplicity is always an advantage, and in some cases it is decisive.

Jorge Mateu directs our attention to the possibility that betting might be the only testing method simple enough to be implemented. Suppose a probability model P is defined in such a way that its probabilities and expected values can be obtained only by simulation; see, for example the spatio-temporal models in Tamayo-Uria et al. (2014). There may be obvious test statistics, but the simulations required to estimate tail probabilities may be impractical, and it may be even more difficult to identify alternatives and make power calculations. But if we choose a bounded nonnegative test statistics T, then estimating $\mathbf{E}_P(T)$ to a couple significant figures may be much easier than estimating a tail probability. This will allow us to make the bet $S := T/\mathbf{E}_P(T)$. We can obtain the implied target with just one more simulation, because $\mathbf{E}_Q(\ln S) = \mathbf{E}_P(S \ln S)$. (Here, as Bruce Levin has pointed out to me, we are simulating Q with importance sampling.) Following the example of Augustine Kong and Nancy J. Cox, as reported by Xiao-Li Meng in his comments, we might then study the parametric model defined by $P_\theta(y) := P(y) \exp(\theta \ln S)/c_\theta$.

Testing is not our only task. I just suggested that we replace the term 'hypothesis testing' with 'prediction testing', because predictions are the only thing we can test. But testing predictions is not the statistician's only task. For one thing, we must make predictions. When we say that we are testing a hypothesis, we are usually constructing a complex argument that involves repeatedly making and testing predictions. Calling this process testing gives it a patina of objectivity that can enhance the statistician's authority but may come back to bite her.

In this iterative process of predicting and testing we deal with an important issue that both Priyantha Wijayatunga and Sander Greenland raise: uncertainty about the assumptions on which predictions are based. This is primarily an issue about how we make predictions, not about how we test them. If we multiply our money a lot betting against the predictions, they are discredited no matter how confident we were in their assumptions. If the predictions withstand many tests by many able and well informed scientists, the assumptions may be better than we thought.

Wijayatunga also raises another important issue: how accurate we need predictions to be. When we are testing with a single bet against a probability distribution (or with a strategy for betting on a sequence of outcomes for which we have a joint probability distribution), the implied alternative gives us an opportunity to answer this question. If the predictions provided by the null and by the implied alternative do not differ enough for us to care, then the study is of no value. When the predictions we are testing are not provided by a comprehensive probability distribution, as in meta-analysis or when we are testing time-varying forecasts, we can instead bet against upper and lower probabilities obtained by expanding each point prediction to an interval that represents the precision that matters (Shafer & Vovk, 2019, Chapter 6). A related idea is to introduce transaction costs, as in Wu and Shafer (2007).

When we are making predictions, playing the role of Forecaster in my Protocol 4, for example some of the issues raised by Stephen Senn arise.

- Forecaster must avoid making offers that open him to arbitrage. This is assured in Protocol 4 by the requirement that his offers be defined by a probability distribution.
- The signal x_n may be thought of as a vector of covariates. Forecaster must decide which of them to use and how.

The statistician must also decide how to use the covariates in choosing her bets or the strategy for betting that she prescribes for Sceptic.

Frank Coolen raises the question of how we design an experiment to test P. This is another of the statistician's many tasks. The notion of implied target, like the notion of power, should help us choose between experiments, but it does not otherwise help us design one.

Coolen also mentions the problem of pooling expert opinion to form probability distributions. This is one way Forecaster can make his predictions. Methodology on pooling opinion developed outside the statistical community, such as the work on prediction with expert advice, should also be part of the statistician's toolbox. Something may be gained by posing the problem in terms of a betting protocol like Protocol 4 (Shafer & Vovk, 2019, Chapter 12). The work by Frank Hampel that Coolen cites is also about how to make probabilistic predictions, not about how to test them.

Several discussants emphasize decision problems. As I said in my paper, I consider decision theory an important chapter in statistical methodology. In particular, I believe there are many situations where costs or utilities are available and the Neyman–Pearson lemma is applicable. In these situations, we are making an accept/reject decision that will not be revisited, and so we want to maximize $Q(S \geq 1/\alpha)$ rather than $\mathbf{E}_Q(\ln S)$. I also believe that there are many situations where we have reasonable ingredients for Bayesian assessments.

Contrary to R. A. Fisher's polemical suggestion that the Neyman–Pearson theory belongs only in industrial settings, we know that it is often applied to good effect in scientific investigations where the abundance of possible choices is so great that a preliminary accept/reject screening is required. Christine P. Chai notes that such screening can also be used to discard variables from a study. Chai characterizes this as a use of p-values, but when we use a cut-off, we are doing Neyman–Pearson (fixed-level) significance testing rather than using a p-value as a means of communication.

Chloe Krakauer and Kenneth Rice propose that we think of significance testing as a decision problem and discuss Bayesian solutions. A salient aspect of their proposal is that they consider three possible decisions: decide that a parameter is negative, decide that it is positive or make no decision. Judging from their tone, I think they would agree with my own first reaction: their proposal may be helpful in some but hardly all cases where statisticians and scientists have been using significance tests. If I am testing the predictions of a theory in which many people are interested, those who come

after me to test its other predictions will be interested in what evidence has been accumulated so far, but not in what decision I made.

Krakauer and Rice conclude with a question to me: 'We welcome Prof Shafer's thoughts on quantifying the plausibility of hunches, and how any corresponding calculus differs from that of Bayes'. Here 'hunch' may refer to my statement that my choice of S and hence Q 'may be guided by some hunch about what might work…' The simplest answer to their question is that S quantifies my hunch but not its plausibility. This is one respect in which my proposal is non-Bayesian.

The role of the implied alternative. Christian Hennig applauds the notion of implied alternative as a tool to understand tests better but finds it too sophisticated for the communication of statistical results. He foresees, no doubt correctly, that it could generate its own abuses, and he fears that the whole betting picture, because bets are made to win, may lead to yet more misleading significant results.

In reflecting about these concerns, we need to recognize that different levels of simplicity and sophistication are needed for different audiences. Betting scores are simple enough to be communicated in newspapers, where they have the advantage that they communicate not only the strength of the evidence but also its uncertainty; everyone is immediately aware that a betting score depends on the choice of bet. The notion of an alternative target is indeed much more sophisticated, but it should have its place in communications among research workers. As my simple examples demonstrate, *bet/implied target* work together in a much simpler and less confusing way than the *significance level/p-value/power* triplet we now try to teach research workers.

The research workers with whom I have interacted the most in recent decades are professors in accounting and finance and doctoral students who aspire to this role. So I am painfully aware of the misuse of the concept of p-value and the non-use of the concept of power that characterizes the "top journals" in these fields. In my paper, I cited Cready (2019), Cready et al. (2019), and Harvey (2017), which provide glimpses into this situation. Here, as in a number of other research fields, the time has surely come to try something different.

Hennig nevertheless worries that I *seem* to imply that scientists should want to win their bets, and that this *seems* to take for granted 'the incentive of journals for finding significance'. Perhaps his use of *seem* acknowledges the possibility that abuses might be mitigated. We do want scientists to try very hard to win their play-money bets, no matter whether the predictions they are testing are their own or others'. But the implied target could actually help us rectify incentives. Its adoption by journals would force authors to evaluate the scientific merit of a proposed study. A low implied target means the study has little merit. A plausible implied alternative and high implied target means the study will be informative *regardless of its outcome*. Journals could provisionally accept such informative studies before they are carried out, thus encouraging the preregistration of studies and lessening the incentive to get statistical significance by hook or by crook.

Morality and objectivity. My friend and colleague Harry Crane wants us to bet real money. Philip Dawid and Christian Hennig worry, in contrast, that the negative consequences of betting make it problematic as a basis for communication. Even though I did bet $2 on a horse a few years ago (at my request, Harry was showing me around the local racetrack), my own attitude towards gambling is rather negative. When proposals to relax laws restricting betting were on the ballot in states where I lived, I always voted no. But a vice is not always curtailed by suppressing understanding of it. By forgetting 19th-century insights into betting systems, we have facilitated their replication in 20th- and 21st-century finance (Crane & Shafer, 2020).

Testing by betting is so natural an idea that its absence from statisticians' discourse for hundreds of years cannot be merely an oversight. Statisticians usually avoid translating their methods and insights back into betting language, even though we all know that probability's rules derive from betting. Is

this avoidance due primarily to moral objections? I suspect that another motive is at play: the pursuit of objectivity. We see this already in the first pages of Jacob Bernoulli's *Ars Conjectandi*, the book that first sought to turn the betting calculus into a general theory of probability. Bernoulli replaces Huygens's betting arguments with reasoning about equally possible cases. The bettor, the subjective constituent of the story, is disappeared.

This motive endures. The statistician's public still wants objective results. But there are now also many ready to poke holes in any scientific claim that has social implications. It is my hope that the language of betting can help us educate the public about why and where choice is needed when testing scientific claims statistically.

Martingales. As soon as it considers a sequence of outcomes, probability theory becomes the theory of martingales. Blaise Pascal, Christiaan Huygens and Abraham De Moivre did not use the word *martingale*, but to find the value of a payoff that depends on multiple successive outcomes, they constructed betting strategies that yield the payoff. By the end of the 19th-century, casino-goers called almost any betting strategy a *martingale*. Jean Ville (1939) used the word instead for a betting strategy's capital process—the sequence of random variables that represents the capital of a bettor following the strategy. Abraham Wald, who was familiar with Ville's work in the 1930s, learned after he came to the United States that mathematical statisticians called a martingale a sequence of likelihood ratios. Joseph Doob adopted Ville's word *martingale* but used only its measure-theoretic definition. The word has now been adopted in a number of branches of statistics, but its betting meaning and fundamental role in probability theory usually remain unspoken. It is a measure of the degree to which we have suppressed the role of betting in probability that some of our discussants can discuss the identity 'martingale = capital process' almost as if it were news.

As I noted in section 4.1 of my paper, a global bet in a stochastic process can be implemented by a strategy for Sceptic's step-by-step betting; see Protocol 2. The betting score is thus the final value of a nonnegative martingale. This idea is generalized in many directions in Shafer and Vovk (2019); there we call capital processes *nonnegative supermartingales*, because we allow Sceptic to make disadvantageous bets, and because betting offers may be one-sided and fragmentary.

Aaditya Ramdas's very rich comments emphasize the relation between martingales and bets in the measure-theoretic case. It will be very interesting to see how and to what extent his results extend to nonnegative supermartingales in the game-theoretic framework.

Tze Leung Lai has made innovative contributions to the use of martingales in statistics for many decades, and I am honoured that he and Anna Choi have contributed to the discussion. I am especially glad that they call attention to the work by Lai, Gross and Shen, which builds on work that Françoise Seillier-Moisewitsch and Philip Dawid published in 1993. Seillier-Moisewitsch and Dawid's basic insight, that a martingale central limit theorem can be based on only what happens on the path taken by a stochastic process, was crucial for the development of game-theoretic probability.

Examples. As Paul Smith reports in his note on the live chat, Peter Grünwald has posted some detailed examples of testing by betting at safestatistics.com. The simple example I discussed in section 2.4 of my paper, that of testing whether a normal mean is zero, is also much more than a bauble. For two centuries, beginning with Laplace and Fourier, statisticians have tested the difference between two proportions using a normal approximation. This was already happening in medicine in the 1830s. So perhaps my simple classical example was already a first step towards meeting Sander Greenland's challenge to show how testing by betting translates into "real medical applications".

The unity and diversity of probability. In the 1980s, I argued that we should understand probability and mathematical statistics in terms of betting games. These games stand at the centre of a wide circle of ideas. Different statistical methods use them in different ways. Different interpretations of probability should not be understood as different ways of assigning meaning to numbers but rather as different

ways of assigning reference to entire games (Shafer, 1990). Beginning in the late 1990s, I learned from Vladimir Vovk how we can formalize betting games in game theory, distinguishing the roles of different players while reproducing and extending the classical mathematics of probability. This has only strengthened my belief that betting games can be used by statisticians, scientists and other analysts in a variety of ways. Some probability arguments merely draw analogies with betting games. Others use a betting game as a model of a physical or social process. Others put a decision maker in the position of one of the players in a betting game. Others use the results of one of more betting games as evidence for incidentally related questions.

In the paper under discussion, I develop one way of using a betting game to test predictions. My passionate exposition of it may have led a few of the discussants to worry that I was arguing against the ways they have been using probability. I hope this response has convinced them otherwise.

In particular, let me assure Kuldeep Kumar that I fully support the use of Neyman–Pearson decision theory in tasks such as classification. The work on conformal prediction to which I have contributed falls into this category (Vovk et al., 2005). Testing by betting is also not in competition with Bayesian inference, as Barbara Osimani's comments may suggest. Bayes's theorem is used to find probabilities for hypotheses, not to test predictions. Jeffreys sought to use Bayesian intuitions in testing, but as we have seen, testing by betting re-interprets rather than rejects many of his tests.

Let me similarly assure Ryan Martin that I have not renounced my own work on Dempster–Shafer theory and that I admire his and Chunhai Liu's refinements of it. Their interpretation of *p*-values as plausibilities is especially revealing. I see Dempster–Shafer theory, however, as only one tool for constructing arguments. Martin and Liu's inferential models are similarly only one tool. There is no single tool that should be used in every analysis or argument that draws on the idea of a betting game.

I would also avoid trying to stuff every insight afforded by one tool into the use of a different tool. Arthur Paul Pedersen and some participants in the live chat ask whether insights about utility should be brought into testing by betting, and Osimani asks whether prior information can be used in testing by betting in the same spirit as it is used in a Bayesian analysis. The ideas advanced by Judith ter Schure may be responsive to these questions. To my mind, more formal efforts in these directions would only muddy the waters. No sense in trying to use a screw driver as a saw.

Conclusion. Having disagreed with some of the discussants on small points, let me thank them all again for their interest, their support and the many gems of insight they have added. Some have even given me valuable feedback on drafts of this response.

I especially appreciate Jorge Mateu's succinct statement of one of the most important virtues of betting scores: the uncertainty associated with a large betting score is highlighted as much as the certainty provided by a small *p*-value is sometimes exaggerated.

REFERENCES

Bienvenu, L., Shafer, G. & Shen A. (2009) On the history of martingales in the study of randomness. *Electronic Journal for History of Probability and Statistics* 5 (1), 1–40.

Breiman, L. (2001) Statistical modeling: The two cultures (with discussion). *Statistical Science* 16(3), 199–231.

Cournot, A.A. (1843) *Exposition de la théorie des chances et des probabilités*. Paris: Hachette. Reprinted in 1984 as Volume I (Bernard Bru, editor) of Cournot (2010).

Cournot, A.A. (1973–2010) *Œuvres complètes*. Paris: Vrin. The volumes are numbered I through XI, but VI and XI are double volumes.

Crane, H. & Shafer, G. (2020) Risk is random: The magic of the d'Alembert. Working Paper 57, Available from: https://www.probabilityandfinance.com.

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

Dawid, A.P., de Rooij, S., Shafer, G., Shen, A., Vereshchagin, N. & Vovk, V.(2011) Insuring against loss of evidence in game-theoretic probability. *Statistics and Probability Letters* 81(1), 157–162.

Fama, E.F. (1991). Efficient capital markets: II. *The Journal of Finance* 46(5), 1575–1617.

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

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

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford.

Kass, R.E. & A.E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association* 90(430), 773–795.

Shafer, G. (1990) The unity and diversity of probability (with discussion). *Statistical Science* 5(4), 435–462.

Shafer, G. (2019a) The language of betting as a strategy for statistical and scientific communication. arXiv:1903.06991 [math.ST].

Shafer, G. (2019b) On the nineteenth century origins of significance testing and p-hacking. Working Paper 55, Available from: https://www.probabilityandfinance.com.

Shafer, G. (2019c) Pascal's and Huygens's game-theoretic foundations for probability. *Sartoniana* 32, 117–145.

Shafer, G. & Vovk, V. (2001) *Probability and finance: It's Only a Game!* New York: Wiley.

Shafer, G. & Vovk, V. (2006) The sources of Kolmogorov's Grundbegriffe. *Statistical Science* 21(1), 70–98.

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

Tamayo-Uria, I., Mateu, J. & Diggle, P.J. (2014) Modelling of the spatio-temporal distribution of rat sightings in an urban environment. *Spatial Statistics* 9, 192–206.

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

Vovk, V. (2019). Non-algorithmic theory of randomness. arXiv:1910.00585 [math.ST].

Vovk, V. & V'yugin, V.V. (1993) On the empirical validity of the Bayesian method. *Journal of the Royal Statistical Society, Series B* 55(1), 253–266.

Vovk, V. & Wang, R. (2019) Combining e-values and p-values. arXiv:191206116v1 [math.ST], to appear in Annals of Statistics as "E-values: Calibration, combination, and applications".

Vovk, V., Gammerman, A. & Shafer, G. (2005) *Algorithmic learning in a random world*. Berlin: Springer.

Waudby-Smith, I. & Ramdas, A. (2020) Variance-adaptive confidence sequences by betting. arXiv:2010.09686 [math.ST].

Wu, W. & Shafer, G. (2007) Testing lead-lag effects under game-theoretic efficient market hypotheses. Working Paper 23, Available from: https://www.probabilityandfinance.com.

How to cite this article: Shafer G. Author's reply to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *J R Stat Soc Series A*. 2021;184:432–478. https://doi.org/10.1111/rssa.12672