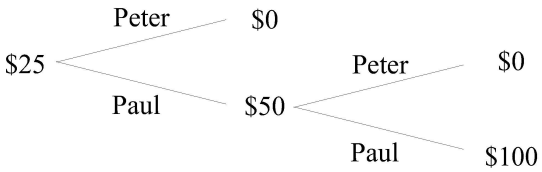
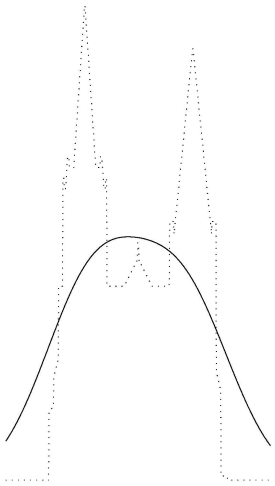
The Language of Betting as a Strategy for Statistical and Scientific Communication

Glenn Shafer, Rutgers University





The Game-Theoretic Probability and Finance Project

Working Paper #54

First posted March 5, 2019. Last revised March 14, 2019.

Project web site:
http://www.probabilityandfinance.com

Abstract

The established language for statistical testing — significance levels, power, and p-values — is overly complicated and deceptively conclusive. Even teachers of statistics and scientists who use statistics misinterpret the results of statistical tests, tending to misstate their meaning and exaggerate their certainty. We can communicate the meaning and limitations of statistical evidence more clearly using the language of betting.

This paper calls attention to a simple betting interpretation of likelihood ratios. This interpretation leads to methods that lend themselves to meta-analysis and accounting for multiple testing. It is closely related to the interpretation of probability as frequency, but it does not encourage the fallacy that probabilistic models imply the existence of unseen alternative worlds.

For more on the betting interpretation of probability, see [42] and the other working papers at `probabilityandfinance.com`.

1 Introduction

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 lead to overconfidence in the results of statistical testing and neglect of relevant information about how the results are obtained. This was already apparent in the 19th century. In 1826, Joseph Fourier published a table of confidence limits for inferences about means from census data, and by 1843 Antoine Augustin Cournot was deploring the widespread practice of announcing such inferences without revealing the search that produced them. In recent decades the problem has become more salient than ever, 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 [47]. 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 [35, 22]. Andrew Gelman and John Carlin, in their commentary on Blakely McShane and David Gal's review of these studies, conclude that the most frequently proposed solutions to this problem (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 [21]. The only solution, they argue, is "to move toward a greater acceptance of uncertainty and embracing of variation" (page 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 is the best evidence we can have that the probabilities are wrong, but everyone understands that such success may be mere luck. Moreover, candor about the betting aspects of scientific exploration can communicate truths about the games scientists must and do play — honest games that are essential to the advancement of knowledge.

A natural way to test probabilities is to make a bet that may multiply the money one risks by a large factor. Section 2 elaborates on the simple but usually overlooked observation that this is equivalent to buying a likelihood ratio for its expected value. Unlike a Neyman-Pearson significance test, such a bet is not all-or-nothing; it can succeed to different degrees. This leads to the notion of the bet's *implied target*, which promises to be more usable than Neyman and Pearson's notion of power.

Testing by betting could improve practice in several ways. First, when we report that a bet has discredited a hypothesis, the remaining uncertainty is glaringly clear. Second, the betting interpretation can give scientists a clearer view

of what a study may accomplish before it is carried out and thereby decrease the number of studies that erroneously detect effects. Third, it is always legitimate to continue betting, and this makes each individual study a more informative element of a research program or a meta-analysis.

In Section 3, we consider statistical modeling and estimation. Here the betting moves partly off-stage. Just as the conventional picture of a statistical model involves partial knowledge of a probability distribution, the betting picture of a statistical model involves seeing only some moves of a betting game. We do not see how a bet pays off. But we can nevertheless equate the model's validity as a description of a phenomenon with the futility of betting against the model, and this can translate into a *warranty* about one or more unknown parameters.

The thesis that a probabilistic model's validity consists of its ability to withstand betting can also be applied to models in statistical mechanics and macroeconomics. Here observation is at a macroscopic level, no individual moves being observed, but the inability of simple betting strategies to multiply their capital substantially implies observable macroscopic regularities. Implications for the interpretation of probability are briefly explored in Section 4.

The paper concludes with two appendices. Appendix A provides translations from French to English of some relevant passages from Fourier and Cournot. Appendix B relates the betting interpretation of probability to the thesis that the meaning of probability lies in its prediction of high-probability events and quotes explanations of this thesis by Abraham Wald and Trygve Haavelmo.

2 Testing a probability distribution by betting

You claim that a probability distribution P describes a certain phenomena 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,[1] a natural way to proceed is to interpret P 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 is the best evidence I can have against P. 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

[1] Here I follow a popular convention introduced by Andrei Markov in 1900 [33]: an unknown quantity is denoted by an upper case Latin letter at the end of the alphabet, and an actual or possible value of the quantity is denoted by the corresponding lower case letter.

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. So this use of the language of betting is not a chapter in decision theory. It requires neither the evaluation of utilities nor any Bayesian reasoning.

The betting may even be hypothetical. But I must declare my hypothetical bet before the outcome y is revealed; one advantage of the betting language is that this requirement is built into the notion of a bet. In some situations, the requirement that the bet be specified in advance can be replaced by the requirement that the bet be relatively simple. The finiteness of our language limits the number of simple bets, and if there is enough data they can be combined into a single bet [5]. But assessing simplicity can be problematic; see the last paragraph of the quotation from Cournot in Appendix A.

The idea of testing probabilities by trying to multiply the money risked is so natural that it must be very old. I am unable, however, to report many instances of its being spelled out. It was implicit, at least, in Jean Ville's critique of Richard von Mises's collectives [45]. But where do we find it in the statistical literature?

Many writers on statistical testing allude to betting from time to time. But the bets mentioned are usually all-or-nothing bets, bets on events rather than payoffs that can take more than two values. And betting never takes center stage, even though it is central to everyone's intuitions about probability. It seems that mathematicians have systematically suppressed the betting aspect of statistical testing to make that testing look more objective — more scientific.

2.1 Betting score = 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 $S \geq 0$, this tells us that 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

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

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

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

Many mathematicians, including R. A. Fisher, have advocated using likelihood as a direct measure of evidence.[2] In Chapter III of his *Statistical Methods and Scientific Inference*, first published in 1956 [18], Fisher suggested that the plausible values of a parameter indexing a class of probability distributions are those for which the likelihood is at least (1/15)th its maximum value. Later authors, including A. W. F. Edwards [15] and A. P. Dempster, have argued at length for using likelihood ratios to measure evidence against null hypotheses. An article on the topic by Dempster, first published with discussion in the proceedings of a 1973 conference, was reprinted in 1997 along with further discussion [14]. 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.

Discussion of betting is similarly conspicuous by its absence from statistical theory that uses the concept of a martingale. When P and Q are probability distributions for a sequence $Y_1, Y_2, \ldots$, the sequence

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

is a nonnegative martingale. Everyone who works with martingales is aware of the betting ancestry of the word "martingale", but Joseph L. Doob purified the word of this heritage when he made it a technical term in modern probability theory. The concept of a martingale is now widely used in sequential analysis, time series, and survival analysis [1, 30], but I have not found in these literatures any use of a nonnegative martingale to score the evidential value of the outcome of a bet.

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 conditions on my choice of S. The choice may be guided by some hunch about what might work, or I may act on a whim. It is only required that I make the choice before seeing Y's value or getting any information about it.

Suppose I do have a hunch that a different probability distribution Q is a better 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

$$\mathbf{E}_Q\left(\ln \frac{Q}{P}\right) \geq \mathbf{E}_Q\left(\ln \frac{R}{P}\right) \tag{1}$$

for any probability distribution R for Y.[3] But why should I choose S to

[2] In 1921 [16], Fisher introduced the name "likelihood" for the function that assigns to each of several probability distributions (or hypothetical populations) the probability it assigns to an observed outcome. But the idea of using this function as a measure of evidence goes back to the 18th century [44, Chapter 16].

[3] Many readers will recognize $\mathbf{E}_Q\left(\ln(Q/P)\right)$ as the Kullback-Leibler divergence between Q and P. In the terminology of Kullback's 1959 book [29, page 5], it is the mean information for discrimination in favor of Q against P per observation from Q.

Table 1: Elements of a study that tests a probability distribution by betting. The proposed study may be considered meritorious and perhaps even publishable when the null hypothesis P and the implied alternative Q are both initially plausible and the implied target is reasonably large. A large betting score 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	$S(y)$

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?[4]

When S is the product of many successive factors, $\mathbf{E}(\ln S)$ measures its rate of growth.[5] So maximizing $\mathbf{E}(\ln S)$ may be appropriate when a proposed study is part of a larger scientific enterprise, where a promising but inconclusive result may lead to further investigation by the same scientists or others.

Suppose, for example, that P is purported to describe $Y_1, Y_2, \ldots$. I first test P by buying $S_1(Y_1)$ for \$1. I obtain a mediocre betting score; $S_1(y_1)$ is greater than 1, but not by much. So I test again, and my second score $S_2(y_2)$ is again mediocre. Consider these two ways we might fill out this story:

1. I made the second bet by taking another \$1 out of wallet. So I risked a total of \$2, and my final betting score is the mediocre $(S_1(y_1) + S_2(y_2))/2$.

2. I made the second bet using the winnings from the first: $\mathbf{E}_P(S_2) = S_1(y_1)$. In this case my final betting score is $S_1(y_1)S_2(y_2)$.

[4] As discussed further in Subsection 2.2, maximizing $Q(S \geq 1/\alpha)$ leads to an all-or-nothing bet corresponding to a Neyman-Pearson test.

[5] The idea of maximizing $\mathbf{E}(\ln S)$ in order to maximize a growth rate was explained succinctly by John L. Kelly, Jr. in 1956 [28]: "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 [6], information theory [12], finance theory [32], and machine learning [8].

The second option is more attractive. So it makes sense for me to aim for a large value of $S_1 S_2$ rather than a large value of $S_1 + S_2$.

The scientist's game

Consider the scientist who is searching for factors that might influence a phenomenon. After exploring 20 different possible factors (20 different chemicals in a medical study or 20 different stimuli in a psychological study), she finds a factor that that has an apparent effect — one that could have happened by chance only 5% of the time. How seriously should we take this apparent discovery?

One simple answer is that the p-value of 5% should be multiplied by 20; this is the Bonferroni adjustment. It has a betting rationale; we can suppose that the scientist has put up \$1 each time she tests a factor, thereby investing a total of \$20. On her 20th try, she multiplied her \$1 by 20, thus obtaining a betting score of 20. When we recognize that she actually invested \$20, not merely \$1, we conclude that her total betting score may be as low as 20/20, or 1.

On the other hand, there is a less severe and more reasonable way to use betting language here. As when the same hypothesis is tested repeatedly, we can suppose that the scientist uses the winnings from the each bet to make the next bet. Thus the initial investment will be only \$1, and the final betting score will become an evaluation of the scientist's broader hypothesis that some factor in the class being studied does make a difference.

In many fields, the increasing resources being devoted to the search for significant effects has led to widespread and justified skepticism about published statistical studies purporting to have discovered such effects. This is true for both experimental studies and studies based on databases. A recent replication of published experimental studies in social and cognitive psychology has shown that many of these studies cannot be replicated [9]. A recent review of database studies in finance has noted that although an immense number of factors affecting stock prices have been identified, few of these results seem to be believed, inasmuch as each study ignores the others [27]. These developments confirm that we need to report statistical results in ways that embed them into broader research programs. Betting scores may help.

Bayes factor?

Suppose I believe that there does exist a probability distribution that is a valid description of Y. I believe this probability distribution is in a known class $\mathcal{C}$ of probability distributions, and I have probabilities that express my beliefs about which distribution in $\mathcal{C}$ it is. Then I can obtain a single probability distribution Q by averaging the distributions in $\mathcal{C}$ with respect to these probabilities. In this case, it is conventional to call the likelihood ratio $P(y)/Q(y)$ a *Bayes factor*.

When I announce S as my bet against P, I am not obliged, however, to believe that there exists any probability distribution that is a valid description of Y. We have seen that $S = Q/P$ for some probability distribution Q, but this implies neither that I consider Q a valid description of the phenomenon nor

that I have beliefs that make $Q(y)/P(y)$ my Bayes factor. I may have chosen S for any reason or for no reason at all. The only condition on my action is that I choose S before I learn anything more about Y.

Moreover, the notion of an alternative hypothesis fades from view when we combine successive tests of P made by different individuals or teams. If P provides probabilities for a phenomenon $Y_1, Y_2, \ldots$, one team may test P with a bet $S_1(Y_1) = Q_1(Y_1)/P(Y_1)$. A second team, after seeing the betting score $S_1(y_1)$, may then test P with a bet $S_2(Y_2) = Q_2(Y_2)/P(Y_2)$. If the combined betting score $S_1(y_1)S_2(y_2)$ is large, then a meta-analyst may conclude that P has been discredited. But this is justified only by the fact that the two teams, acting in succession, bet against P in a way that multiplied the money risked by a large factor. The did not do so by making a single bet $S(Y_1, Y_2)$ at the outset. So there is no comprehensive alternative probability distribution Q in view. There is not even a team or individual in view to whom we can attribute subjective probabilities Q_1 for Y_1 and subjective probabilities Q_2 for Y_2 given $Y_1 = y_1$.

Bayesian interpretation?

I may have previous evidence against P. I might, at least in personal and private deliberation, consider this previous evidence comparable to having obtained some particular betting score S_0 from a previous bet against P. Then I can treat the product $S_0S(y)$ as my total score against P. This process resembles a Bayesian analysis, where prior odds are multiplied by a likelihood ratio to obtain posterior odds. But the logic is not Bayesian and the conclusion is not Bayesian. The score $S_0S(y)$ is understood as the payoff achieved by a bet that had many possible payoffs, not as odds for an all-or-nothing bet whose payoff depends on finding out whether P is valid or not.

2.2 Neyman-Pearson tests as all-or-nothing bets

Although statistical testing can be traced back to the 17th century, the abstract formulation we now teach is usually attributed to Jerzy Neyman and E. S. Pearson's celebrated 1928 article in *Biometrika* [39].

In Neyman and Pearson's theory, we specify a *significance level* $\alpha \in (0, 1)$, usually very small, and a set E of possible values of Y such that $P(Y \in E) = \alpha$. We reject P if the actual value y is in E. This can be expressed in betting terms: we pay \$1 for the bet S_E defined by

$$S_E(y) := \begin{cases} \frac{1}{\alpha} & \text{if } y \in E \\ 0 & \text{if } y \notin E; \end{cases} \tag{2}$$

when E happens, we have multiplied the money we risked by the large factor $1/\alpha$, and this discredits P. Let us call a bet of the form (2) an *all-or-nothing* bet.

Neyman and Pearson suggested that for a given significance level α, we chose 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.[6] 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

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

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. 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 outcomes that reject P.

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. But as we will see, the success of the Neyman-Pearson bet in such cases is not always convincing evidence against P.

2.3 Power and implied targets

In the Neyman-Pearson theory, the probability $Q(\text{test rejects } P)$ is called the *power* of a significance test of P against Q. For nearly a century, mathematical statisticians have decried the use of tests with low power, but the question of a test's power is often ignored by scientists. There are obvious reasons for this. A scientist may find it impossible to obtain enough observations to provide high power under a reasonable alternative Q, and it may be difficult to say whether an alternative is reasonable or unreasonable.

A more fundamental difficulty with the notion of power is that refers to a different game than the one the scientist is playing. In Neyman and Pearson's game, a player chooses "reject P" or "accept P". This is appropriate if we are testing a widget that is to be put on sale if accepted or returned to the factory for rework if rejected, never in either case to be tested again. But the scientist is looking for evidence against a hypothesis P that may be tested again many times in many ways. Fisher famously criticized Neyman and Pearson for confusing the scientific enterprise with the problem of "making decisions in an acceptance procedure" [18, Chapter 4]. I propose that we build on Fisher's critique by considering games that are more appropriate as models for the scientific enterprise.

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 $\mathbf{E}_Q(\ln S)$. So the scientist might hope for a betting score whose logarithm is in the ballpark of $\mathbf{E}_Q(\ln S)$ — i.e., a betting score that is in the ballpark of

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

[6] Asking the reader's indulgence, I leave aside the fact that the discreteness of P and Q can make it impossible to do this precisely or uniquely.

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

The notion of an implied target is the natural replacement for the notion of power when we consider bets that are not necessarily all-or-nothing bets. One of its advantages is that the scientist cannot avoid defining it by refusing to specify a particular alternative. The implied alternative Q and the implied target $\mathbf{E}_Q(\ln S)$ are determined as soon as the null hypothesis P and the bet S are specified, and 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 the standard way of testing probability distributions, the implied target will inevitably be provided by the software that implements such tests, and editors will inevitably demand that it be included in any publication of results. See Table 1.

2.4 Two examples

The following examples are instructive. In both examples, P and Q are continuous and so the likelihood ratio is the ratio of their densities, $q(y)/p(y)$. In both examples, the Neyman-Pearson test rejects P, but the likelihood ratio favors Q only slightly (Example 1) or favors P (Example 2).

Example 1 is a typical example of a study that should not be implemented because the only plausible alternatives lead to low implied targets and Neyman-Pearson tests with low power. In Example 2 the alternative Q leads to an astronomical implied target and a Neyman-Pearson test with high power, but rejection by the Neyman-Pearson test is not necessarily evidence in favor of Q.

Example 1. In their discussion of the p-value communication problem [21, page 900], Gelman and Carlin comment on the notion of practical significance:

> ... the distinction between practical and statistical significance does *not* resolve the difficulties with p-values. The problem is not so much with large samples and tiny but precisely measured effects but rather with the opposite: large effect-size estimates that are hopelessly contaminated with noise. Consider an estimate of 30 with standard error 10, of an underlying effect that cannot realistically be much larger than 1. In this case, the estimate is statistically significant and also practically significant but is essentially entirely the product of noise. This problem is central to the recent replication crisis in science ... but is not at all touched by concerns of practical significance.

We may elaborate their example as follows.

- 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.
- Statistician A uses the Neyman-Pearson bet with $\alpha = 0.0015$, which rejects P when $y > 29.68$. This test has power of about 6% against Q.
- Statistician B uses the likelihood ratio

$$\frac{q(y)}{p(y)} = \exp\left(\frac{2y-1}{200}\right).$$

  The implied target is $\exp(0.095) \approx 1.10$.
- We observe $y = 30$. Statistician A multiplied the money she risked by $1/0.0015 \approx 666$, while Statistician B multiplied the money she risked by $\exp(59/200) \approx 1.34$.

Example 2. In [14, pages 249–250], Dempster called attention to artificial examples where y is less likely under Q than under P and yet still extreme enough that the Neyman-Pearson test rejects P. Here is one such example:

- P is uniform on $[0, 1]$; $p(y) = 1$ for $y \in [0, 1]$.
- Q has density $q(y) = 121y^{120}$ for $y \in [0, 1]$.
- Statistician A uses the Neyman-Pearson bet with $\alpha = 0.05$, which rejects when $y \geq 0.95$. This test has power of about 99.8% against Q.
- Statistician B uses the likelihood ratio

$$\frac{q(y)}{p(y)} = 121y^{120}.$$

  The implied target is

$$121 \exp\left(\frac{121 \times 120}{122}\right) \approx 5.9 \times 10^{53}.$$

- We observe $y = 0.95$. Statistician A multiplied the money she risked by $1/0.05 = 20$, while Statistician B multiplied the money she risked by $121(0.95)^{120} \approx 0.25$.

Because it gives more and better information about what the scientist may hope to accomplish, the notion of an implied target may prove more useful in practice than the notion of power. If a bet S against P has a high implied target, and the implied alternative Q is plausible before the study is conducted, then the study will be informative, perhaps even publishable, even if the betting score $S(y)$ comes out low.

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

p-value	$\frac{1}{\text{p-value}}$	f(p-value)
0.10	10	0
0.05	20	8.9
0.01	100	20
0.005	200	28
0.001	1000	63

2.5 p-values

In practice, an all-or-nothing test is usually based on a test statistic $T(Y)$; the rejection set E consists of the values y for which $T(y)$ exceeds a given level.

It is conventional to call the quantity

$$p(y) := P(T(Y) \geq T(y)) \tag{3}$$

the *p-value* resulting from the test statistic $T(Y)$ and the outcome y.[7] The p-value measures the evidence against P, because the level-α test based on T will reject if and only if $p(y) \leq \alpha$.

We cannot interpret the quantity $1/p(y)$ as a betting score, because we can bet at the level $p(y)$ only after we know y. Pretending after we see y that we had made a level-$p(y)$ Neyman-Pearson bet would be cheating.

If the test statistic T is specified in advance but a level α is not, then we can bet legitimately on the p-value (3) being small by choosing a decreasing nonnegative function f such that $f(p(Y))$ has expected value 1 or less under P and making the bet S given by $S(y) = f(p(y))$. My favorite f, because it is easy to remember and calculate, is

$$f(p) := \begin{cases} \frac{2}{\sqrt{p}} & \text{if } p \leq \frac{1}{16} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

(To verify that $\mathbf{E}_P(f(p(Y))) \leq 1$, use the inequality $P(p(Y) \leq p) \leq p$.) As Table 2 shows, this turns a p-value of 0.01, which suggests a bet that has multiplied the money it risked by 100, into a genuine betting score of 20.

The alternative distribution Q defined by (4) cannot be considered reasonable, since it gives probability zero to all y for which $p(y) > 1/16$. But the bet provides a quick rule of thumb for someone who hears a p-value quoted and has no other information about the study that produced it.

[7] This formalizes the way early 19th-century statisticians used the table provided by Fourier in the passage quoted in Appendix A. Fourier's ∂, the average observation divided by Fourier's g, is the statistic T, and Fourier's P is the p-value. Later 19th-century and early 20th-century authors, including R. A. Fisher, used the phrase "the value of P" when writing about statistical evidence. This became "p-value" only beginning in the 1970s.

3 Betting games as statistical models

The problem of testing a single probability distribution is only the starting point of statistical theory. To deploy betting language for statistical estimation, we need a more careful account of the betting, and this requires game theory.

The preceding section began with the assertion that a probability distribution describes a certain phenomenon and showed how to test the assertion by betting. Translating this into game theory means giving a protocol that specifies how betting proceeds: how betting offers are made, who decides what offers to accept, and who decides the outcomes. It is then the protocol, not a probability distribution, that represents the phenomenon.

According to R. A. Fisher, the theory of statistical estimation begins with the assumption that the statistician has only partial knowledge of a probability distribution describing a phenomenon.[8] She knows only that it is in a known class $(P_\theta)_{\theta\in\Theta}$. The corresponding game-theoretic assumption is that the statistician stands outside a betting protocol, perhaps instructing the bettor to play a certain strategy but seeing only some of the moves. The parameter θ is one of the moves the statistician does not see. She also does not see how bets come out. But if she believes that the protocol is a valid description of the phenomenon and has no reason to think that a betting strategy for Skeptic that she has specified 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 probability statements made by Fourier in 1826 and the *confidence intervals* defined by Jerzy Neyman in the 1930s [37].

3.1 A protocol for betting at even odds

To see how a probability distribution can be expressed in terms of a betting protocol, consider the simplest example: repeated independent trials of an event that always happens with probability 1/2.

Skeptic bets, and Reality decides. Write $\mathcal{K}$ for Skeptic's capital, and suppose there are 100 trials. Here is the protocol:

Fair Coin Protocol

$\mathcal{K}_0 = 1$.
FOR $n = 1, \dots, 100$:
  Skeptic announces $M_n \in \mathbb{R}$.
  Reality announces $y_n \in \{0, 1\}$.
  $\mathcal{K}_n := \mathcal{K}_{n-1} + M_n(y_n - \frac{1}{2})$.

Like all the protocols we consider, this is a perfect-information protocol; the players move in the order shown, and each sees the other's moves as they are

[8] The idea is much older [25], but Fisher put it in enduring form in 1922 [17]. He distinguished *parameters*, which characterize an unknown probability distribution, from quantities calculated from the data, which he called *statistics*. He systematically used different symbols for the two types of quantities, and he was very proud of the clarity brought by this terminological and notational innovation ([4], p. 81; [26]).

made. We call M_n the *total stakes* for the nth round. To bet on $y_n = 1$, Skeptic chooses $M_n > 0$; then he gains $M_n/2$ if $y_n = 1$ and loses $M_n/2$ if $y_n = 0$. To bet on $y_n = 0$, he chooses $M_n < 0$.

The protocol becomes a game when we add a rule for who wins. We know by Zermelo's theorem that one of the players will have a winning strategy in such a game.

Here is one rule: Skeptic wins if all $\mathcal{K}_n \geq 0$ and either

- $\mathcal{K}_{100} \geq 20$ or
- $\left|\overline{y}_{100} - \frac{1}{2}\right| \leq 0.1$,

where $\overline{y}_{100}$ is the frequency of 1s: $\overline{y}_{100} = \sum_{n=1}^{100} y_n/100$. Otherwise Reality wins.

Skeptic has the winning strategy in this game, a strategy that risks no more than his initial capital $\mathcal{K}_0$ (because it guarantees $\mathcal{K}_n \geq 0$ for all n) and multiplies $\mathcal{K}_0$ by 20 or more unless the frequency of 1s is within 0.1 of 1/2. This is a theorem in game theory, a special case of a game-theoretic central limit theorem [42, Section 2.3]. Perhaps it justifies calling 1/2 a frequency. But we put the protocol to work as the description of a phenomenon not by finding a frequency equal to 1/2 in the phenomenon but by presuming that Skeptic will not multiply the capital he risks by a large factor. When Skeptic plays the strategy just mentioned, he is paying his initial capital $\mathcal{K}_0 = 1$ for his final capital $\mathcal{K}_{100}$, which is a function S of $y = (y_1, \ldots, y_{100})$. If Reality violates $\left|\overline{y}_{100} - \frac{1}{2}\right| \leq 0.1$, then Skeptic's betting score $S(y)$ will be 20 or more, discrediting to this extent the protocol.

Similar games can be used to establish game-theoretic forms of other classical theorems (laws of large numbers, law of the iterated logarithm, Hoeffding's inequality, etc.) for any probability distribution for a sequence $y_1, y_2, \ldots$ [42].

3.2 A protocol for bounded errors

Many payoffs are priced by a probability distribution for Y — all the payoffs $S(Y)$ for which the expected value exists. When we work with games in which betting offers themselves take center stage, it no longer seems natural or necessary for the offers to be so numerous and comprehensive. Some payoffs $S(Y)$ might be priced for sale and others not.[9]

To illustrate this point, consider the following protocol, which represents in a minimal way assumptions about errors of measurement suggested by Carl Friedrich Gauss in 1821 [43]: each error is bounded, and an error of a given size might just as well be negative as positive.

Bounded Error Protocol

$\mathcal{K}_0 := 1$.
FOR $n = 1, \ldots, 100$:
    Skeptic announces $M_n \in \mathbb{R}$.

[9] Limited betting offers have also been studied under the rubric *imprecise probability* [2].

Reality announces $\epsilon_n \in [-1, 1]$.
$\mathcal{K}_n := \mathcal{K}_{n-1} + M_n\epsilon_n$.

On the nth round, Skeptic can buy any multiple (positive, negative, or zero) of the error ϵ_n at the price 0.

Consider this rule for who wins: Skeptic wins if all $\mathcal{K}_n \geq 0$ and either $\mathcal{K}_{100} \geq 20$ or $|\overline{\epsilon}_{100}| \leq 0.272$, where $\overline{\epsilon}_{100} = \sum_{n=1}^{100} \epsilon_n/100$. Otherwise Reality wins. Section 3.3 of [42] uses a game-theoretic form of Hoeffding's inequality to show that Skeptic has a winning strategy in the game defined by this rule. This is a law of large numbers without a probability distribution.

When Skeptic plays a wining strategy in this game, we can say that he is making a bet S; he is paying 1 for a nonnegative payoff $S(\epsilon_1, \ldots, \epsilon_{100}) = \mathcal{K}_{100}$. If $|\overline{\epsilon}_{100}| > 0.272$, so that $S(\epsilon_1, \ldots, \epsilon_{100}) \geq 20$, we can say that Skeptic has discredited the protocol. But because the protocol does not define a probability distribution P, the bet S does not imply an alternative Q and is not a likelihood ratio.

The bound $\left|\overline{y}_{100} - \frac{1}{2}\right| \leq 0.1$ that we obtained for the Fair Coin Protocol is a familiar 95%-probability interval based on the normal approximation to the binomial distribution. Here, in the Bounded Error Protocol, the range for the outcomes is doubled, $[0, 1]$ being one unit in length and $[-1, 1]$ being two units in length. This doubling can account for the bound increasing from 0.1 to 0.2. The additional increase from 0.2 to 0.272 can be attributed to the fact that Skeptic has fewer bets available than when each error is assigned a probability distribution with mean zero.

3.3 When the statistician stands outside the protocol

Let us elaborate the Bounded Error Protocol by supposing that the errors are added to a quantity μ that is being measured:

Measurement Protocol

$\mathcal{K}_0 := 1$.
Reality announces $\mu \in \mathbb{R}$.
FOR $n = 1, 2, \ldots, 100$:
  Skeptic announces $M_n \in \mathbb{R}$.
  Reality announces $\epsilon_n \in [-1, 1]$ and sets $y_n := \mu + \epsilon_n$.
  $\mathcal{K}_n := \mathcal{K}_{n-1} + M_n\epsilon_n$.

As always, this is a perfect-information protocol. In particular, Skeptic sees Reality's moves μ and $\epsilon_1, \ldots, \epsilon_{100}$ as they are made.

We now assume that the statistician is not Skeptic. She sees only part of what happens. She sees the measurements $y_1, \ldots, y_{100}$ but not the quantity μ being measured or the errors $\epsilon_1, \ldots, \epsilon_{100}$. This does not negate the fact that Skeptic has a strategy that multiplies his money by 20 unless $|\overline{\epsilon}_{100}| \leq 0.272$ — i.e., unless μ is in the interval $\overline{y}_{100} \pm 0.272$.

Assuming that the statistician does not see μ, even at the end of the game, she must leave it to Skeptic to test the Measurement Protocol. But if she considers this protocol a valid description of the measurement being made, perhaps

on the basis of other tests when the apparatus involved was used to measure known quantities, she may believe that the strategy for Skeptic we are discussing has not multiplied the capital it risked by a large factor. In this sense, she may claim that Skeptic has provided a *20-fold warranty* that μ is in $\overline{y}_{100} \pm 0.272$. This is another example of our betting language. The 20-fold warranty does not provide any level of irrefutable confidence that μ is in $\overline{y}_{100} \pm 0.272$. It merely asserts that Skeptic has multiplied his money by 20 otherwise. If the statistician sees other evidence that μ is not in the interval, then she may conclude that Skeptic actually has been this lucky.[10]

The scientist will know the betting score that Skeptic has achieved only as a function of μ. But a meta-analyst, imagining that Skeptic has used his winnings from each study in the next study, can multiply the functions of μ to obtain warranties about μ that may be more informative than those from the individual studies.

The particular strategy for Skeptic we are discussing, as described in [42, Section 3.3], can continue for any number of rounds and gives a 20-fold warranty that μ is in $\overline{y}_n \pm \sqrt{(2 \ln 40)/n}$ after n rounds. The statistician can also ask Skeptic for K-fold warranties for values of K other than 20; we can ask what intervals this particular strategy warrants at the K-fold level, and we can also look for other strategies that optimize what can be achieved with K-fold warranty.

The notion of K-fold warranty can be compared with probability statements made by Fourier and with the formal definition of $(1-\alpha)$-confidence given by Neyman [37]. Cournot gave Fourier's probabilities a frequency interpretation by imagining a statistician making the same bet many times in "perfectly similar" situations (see again Appendix A). Neyman extended this picture by arguing that a statistician who makes many $(1-\alpha)$-confidence statements using valid models will be right $(1-\alpha)$ of the time even if the situations and models vary [38]. Cournot's and Neyman's pictures can be extended to K-fold warranties. Skeptic's multiplying his capital by K or more has game-theoretic upper probability $1/K$ or less, and when suitable repetitions are available we can use a game-theoretic law of large numbers to make frequency statements [42]. But none of this is needed; our betting language is enough to explain the meaning of a K-fold warranty, and talk about frequencies can only serve to mislead by making the meaning appear more objective and definitive.[11]

3.4 Probability forecasting with signals

Many betting protocols include a third player, Forecaster, who announces probabilities or otherwise offers bets. In the following example, Forecaster gives a

[10] See [20] for examples of outcomes that cast doubt on confidence statements and would also cast doubt on warranties.

[11] Both warranty statements and confidence statements can be nested. In principle, a given bet by Skeptic determines K-warranties for all $K > 0$; it gives a K-warranty to the set of parameter values for which the moves the statistician sees imply a payoff for Skeptic less than or equal to K, and this set grows as K increases. Confidence sets for different levels α can also be nested, but the different sets are based on different tests [13, 48].

probability on each round using a signal provided by Reality.

Probability Forecasting Protocol

$\mathcal{K}_0 = 1$.
FOR $n = 1, \ldots, 100$:
  Reality announces signal x_n.
  Forecaster announces $p_n \in \{0, 1\}$.
  Skeptic announces $M_n \in \mathbb{R}$.
  Reality announces $y_n \in \{0, 1\}$.
  $\mathcal{K}_n := \mathcal{K}_{n-1} + M_n(y_n - p_n)$.

Forecaster might be a theory or an algorithm that the other players know in advance, or he might be person, such as a weather forecaster, who decides on his predictions in the course of play.[12] In either case, Skeptic has a strategy for testing the probabilities $p_1, \ldots, p_n$ that guarantees that $\mathcal{K}_n \geq 0$ for all n and either $\mathcal{K}_{100} \geq 20$ or $|\overline{y}_{100} - \overline{p}_{100}| \leq 0.272$. This means that Forecaster and Reality must obey a law of large numbers in order to keep Skeptic from discrediting the protocol by multiplying the capital he risks by a large factor. Other strategies for Skeptic enforce other properties that mathematicians expect when $p_1, \ldots, p_n$ are probabilities.[13]

When Forecaster follows a known strategy that depends on the xs, Skeptic's bets translate into predictions about the ys using the xs, and a strategy for Skeptic can provide K-warranties for parameters in a presumed relation between the ys and the xs. This feature of game-theoretic probability is relevant to any field of science or engineering where probabilities are determined by decisions made in the course of an experiment. The examples of least squares estimation, parametric estimation, and quantum mechanics are discussed in [42, Chapter 10].

4 Probability without multiple worlds

Since the 1970s, significance tests and confidence statements have been called *frequentist* and contrasted with Bayesian methods. The adjective refers to the interpretation of probabilities as frequencies. Statisticians who use Bayesian methods may interpret the probabilities given by a statistical model $(P_\theta)_{\theta \in \Theta}$ as frequencies, but in order to apply Bayes's theorem they complement these probabilities with subjective probabilities for θ.

Richard von Mises, A. P. Dempster, and Ian Hacking have occasionally used the adjective *Bernoullian* instead of *frequentist* (see [40] for references). This recognizes that Bernoulli was the first to introduce a non-Bayesian method of statistical estimation, just as Bayes was the first to use Bayes's rule. I favor this use of Bernoulli's name, because it allows us to contrast Bernoullian and

[12] If we suppress the signals and put in Forecaster's place a theory that always sets p_n equal to 1/2, then the protocol reduces to the Fair Coin Protocol.

[13] We can also construct strategies for Forecaster that will defeat most of Skeptic's testing strategies [42, Chapter 12].

Bayesian methods without asserting anything about how the probabilities involved are to be interpreted.

As we have seen, statements about frequencies are among the consequences of the validity of a betting protocol, but they are not fundamental. The fundamental meaning of the validity of a betting protocol or the reliability of its Forecaster is that a strategy for Skeptic will not multiply the capital it risks by a large factor.

The notion that a phenomenological interpretation of probability must be based on frequencies leads to confusion when it is combined with the fact many of the phenomena that we model with probabilities, such as a nation's economy or even (in cosmology) the evolution of the universe, are not repeated. To interpret these probabilities as frequencies, we apparently need to imagine multiple unseen worlds. Most physicists and philosophers reject this conclusion, but the trend towards realism in philosophy in recent decades has given it proponents and made it seem hard to avoid.

In this context, it is instructive that a betting protocol with outcomes $y_1, y_2, \ldots$ defines a probability distribution for $Y_1, Y_2, \ldots$ only in the relatively simple case where there are no signals and no Forecaster. When there is no probability distribution, we are less likely to stumble into the fallacy that we need frequencies in order to interpret a probability distribution.

5 Conclusion

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 *K-fold warranty* (an alternative to $(1-\alpha)$-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.

Appendices

A p-hacking in the 19th century

The practice of searching for statistically significant results and reporting only the ones found is now often called *p-hacking*. Some authors have attributed its current prevalence to shortcomings in Fisher's work or that of Neyman and Pearson.[14] In fact, the use and abuse of p-values emerged as soon as statisticians understood the normal approximation to the binomial distribution and had interesting data to which to apply it.

[14] See, for example, [22] and [23]. For a thorough look at the interaction between Fisher and Neyman and the evolution of their thought, see [31].

In the 1820s, the mathematician Joseph Fourier held a leadership post in the census bureau of the Paris region, and in this role he wrote a manual, published as part of a census report for 1826, for using the probability calculus to interpret census results. This manual included what we might now call a table of significance levels, along with instructions for how to use them to obtain what we now call confidence limits.

Fourier's limits were based on the normal approximation to the probability distribution of the average $\overline{y}$ of independent observations $y_1, \ldots, y_m$. His table used the quantity

$$g = \sqrt{\frac{2}{m}\left(\frac{\sum_{i=1}^{m} y_i^2}{m} - \overline{y}^2\right)},$$

which differs from what we now call $\overline{y}$'s standard error mainly by its inclusion of the factor $\sqrt{2}$. Fourier's table may look more familiar when we add a third column translating units of g into number of standard errors, a standard error being $g/\sqrt{2}$:

units of g	P	units of $g/\sqrt{2}$
0.47708	$\frac{1}{2}$	0.67
1.38591	$\frac{1}{20}$	1.96
1.98495	$\frac{1}{200}$	2.81
2.46130	$\frac{1}{2000}$	3.48
2.86783	$\frac{1}{20000}$	4.06

Fourier wrote that it is "a 19 out of 20 bet" that the error will not exceed $1.38591g$.[15] This is the familiar 95% confidence interval obtained using 1.96 standard errors.

Here, translated from the French, is a passage from Fourier's manual containing the table and a passage from Antoine Augustin Cournot's 1843 book, which shows that statisticians had promptly used Fourier's guidance to p-hack the census results.

A.1 Fourier 1826 [19, pages xxi–xxii]

To complete this discussion, we must find the probability that H, the quantity sought, is between proposed limits A + D and A + D. Here A is the average result we have found, H is the fixed value that an infinite number of observations would give, and D is a proposed quantity that we add to or subtract from the value A. The following table gives the probability P of a positive or negative error greater than D; this quantity D is the product of g and a proposed factor ∂.

[15] Similar betting language had been used earlier by Laplace [7, Volume 2, page 462].

∂	P
0.47708	$\frac{1}{2}$
1.38591	$\frac{1}{20}$
1.98495	$\frac{1}{200}$
2.46130	$\frac{1}{2000}$
2.86783	$\frac{1}{20000}$

Each number in the P column tells the probability that the exact value H, the object of the research, is between $\mathrm{A} + g\partial$ and $\mathrm{A} - g\partial$. Here A is the average result of a large number m of particular values $a, b, c, d, \ldots, n$, ∂ is a given factor, g is the square root of the quotient found by dividing by m twice the difference between the average of the squares $a^2, b^2, c^2, d^2, \ldots, n^2$ and the square A^2 of the average result. We see from the table that the probability of an error greater than the product of g and 0.47708, i.e. greater than about half of g, is $\frac{1}{2}$. It is a 50–50 or 1 out of 2 bet that the error committed will not exceed the product of g and 0.47708, and we can bet just as much that the error will exceed this product.

The probability of an error greater than the product of g and 1.38591 is much less; it is only $\frac{1}{20}$. It is a 19 out of 20 bet that the error of the average result will not exceed this second product.

The probability of an even greater error becomes extremely small as the factor ∂ increases. It is only $\frac{1}{200}$ when ∂ approaches 2. The probability then falls below $\frac{1}{2000}$. Finally one can bet much more than twenty thousand to one that the error of the average result will be less than triple the value found for g. So in the example cited in Article VI, where the average result was 6, one can consider it certain that the value 6 is not wrong by a quantity three times the fraction 0.082 that the rule gave for the value of g.

The quantity sought, H, is therefore between $6 - 0.246$ and $6 + 0.246$.

A.2 Cournot 1843 [10, Section 111]

. . . Clearly nothing limits the number of the aspects under which we can consider the natural and social facts to which statistical research is applied nor, consequently, the number of variables according to which we can distribute them into different groups or distinct categories. Suppose, for example, that we want to determine, on the basis of a large number of observations collected in a country like France, the chance of a masculine birth. We know that in general it exceeds 1/2. We can first distinguish between legitimate births and those outside marriage, and as we will find, with large numbers of observations, a very appreciable difference between the values of the ratio of masculine births to total births, depending on whether the births are legitimate or illegitimate, we will conclude with very high probability that the chance of a masculine birth in the category of legitimate births is appreciably higher than the chance of the event in the category of births outside marriage. We can further distinguish between births in the countryside and births in the city, and we will arrive at a similar

conclusion. These two classifications come to mind so naturally that they have been an object for examination for all statisticians.

Now it is clear that we could also classify births according to their order in the family, according to the age, profession, wealth, and religion of the parents; that we could distinguish first marriages from second marriages, births in one season of the year from those in another; in a word, that we could draw from a host of circumstances incidental to the fact of the birth, of which there are indefinitely many, producing just as many groupings into categories. It is likewise obvious that as the number of groupings thus grows without limit, it is more and more likely *a priori* that merely as a result of chance at least one of the groupings will produce, for the ratio of the number of masculine births to the total number of births, values appreciably different in the two distinct categories. Consequently, as we have already explained, for a statistician who undertakes a thorough investigation, the probability of a deviation of given size not being attributable to chance will have very different values depending on whether he has tried more or fewer groupings before coming upon the observed deviation. As we are always assuming that he is using a large number of observations, this probability will nevertheless have an objective value in each system of groupings tried, inasmuch as it will be proportional to the number of bets that the experimenter would surely win if he repeated the same bet many times, always after trying just as many perfectly similar groupings, providing also that we had an infallible *criterium* for distinguishing the cases where he is wrong from those where he is right.

But usually the groupings that the experimenter went through leave no trace; the public only sees the result that seemed to merit being brought to its attention. Consequently, an individual unacquainted with the system of groupings that preceded the result will have absolutely no fixed rule for betting on whether the result can be attributed to chance. There is no way to give an approximate value to the ratio of erroneous to total judgments a rule would produce, even supposing that a very large number of similar judgments were made in identical circumstances. In a word, for an individual unacquainted with the groupings tried before the deviation δ was obtained, the probability corresponding to that deviation, which we have called Π, loses all objective substance and will necessarily carry varying significance for a given magnitude of the deviation, depending on what notion the individual has about the *intrinsic importance* of the variable that served as the basis for the corresponding grouping into categories.

B Cournot's principle in the 20th century

When data is plentiful and informative, and our protocol defines a probability distribution for the phenomenon it describes, our conclusions will likely not be affected if we substitute for a bet S with a high implied target a bet on the event $S \geq 1/\alpha$ for a small significance level α. This reduces the presumption that a bet against P will not produce a large betting score to the presumption that an event to which P gives high probability will not happen.

The principle that a mathematical probability near one can be interpreted as practical certainty goes back to Jacob Bernoulli. *Cournot's principle* says that this is the only way that mathematical probabilities can be used to describe phenomena. This principle seems to have first been stated by Cournot in his 1843 book, and it was advocated by prominent mathematicians in the mid-20th century [41]. Here I quote two of them: Abraham Wald, who addressed the issue in a lecture at Notre Dame in 1941, and Trygve Haavelmo, who addressed it in a 1944 article that is often seen as the founding charter of modern econometrics [36]. Haavelmo explained that a probability law for a time series (i.e., a stochastic process) can be tested based on one observation if it makes predictions with very high probability about that one observation.

B.1 Wald 1942 [46, pages 1–2]

The purpose of statistics, like that of geometry or physics, is to describe certain real phenomena. The objects of the real world can never be described in such a complete and exact way that they could form the basis of an exact theory. We have to replace them by some idealized objects, defined explicitly or implicitly by a system of axioms. For instance, in geometry we define the basic notions "point," "straight line," and "plane" implicitly by a system of axioms. They take the place of empirical points, straight lines, and planes which are not capable of definition. In order to apply the theory to real phenomena, we need some rules for establishing the correspondence between the idealized objects of the theory and those of the real world. These rules will always be somewhat vague and can never form part of the theory itself.

The purpose of statistics is to describe certain aspects of mass phenomena and repetitive events. The fundamental notion used is that of "probability." In the theory it is defined either explicitly or implicitly by a system of axioms. For instance, Mises defines the probability of of an event as the limit of the relative frequency of this event in an infinite sequence of trials satisfying certain conditions. This is an explicit definition of probability. Kolmogoroff defines probability as a set function which satisfies a certain system of axioms. These idealized mathematical definitions are related to the applications of the theory by translating the statement "the event E has the probability p" into the statement "the relative frequency of the event E in a long sequence of trials is approximately equal to p." This translation of a theoretical statement into an empirical statement is necessarily somewhat vague, for we have said nothing about the meanings of the words "long" or "approximately." But such vagueness is always associated with the application of theory to real phenomena.

It should be remarked that instead of the above translation of the word "probability" it is satisfactory to use the following somewhat simpler one: "The event E has a probability near to one" is translated into "it is practically certain that the event E will occur in a single trial." In fact, if an event E has the probability p then, according to a theorem of Bernoulli, the probability that the relative frequency of E in a sequence of trials will be in a small neighborhood of p is arbitrarily near to 1 for a sufficiently long sequence of trials. If we translate

the expression "probability near 1" into "practical certainty," we obtain the statement "it is practically certain that the relative frequency of E in a long sequence of trials will be in a small neighborhood of p."[16]

B.2 Haavelmo 1944 [24, pages 477–478]

The reluctance among economists to accept probability models as a basis for economic research has, it seems, been founded upon a very narrow concept of probability and random variables. Probability schemes, it is held, apply only to such phenomena as lottery drawings, or, at best, to those series of observations where each observation may be considered as an independent drawing from one and the same 'population'. From this point of view it has been argued, e.g., that most economic time series do not conform well to any probability model, 'because the successive observations are not independent'. But it is *not* necessary that the observations should be independent and that they should all follow the same one-dimensional probability law. It is sufficient to assume that the *whole set* of, say n, observations may be considered as *one* observation of n variables (or a 'sample point') following an n-dimensional *joint* probability law, the 'existence' of which may be purely hypothetical. Then, one can test hypotheses regarding this joint probability law, and draw inferences as to its possible form, by means of *one* sample point (in n dimensions). Modern statistical theory has made progress in solving such problems of statistical inference.

. . .

The class of scientific statements that can be expressed in probability terms is enormous. In fact, this class contains all the 'laws' that have, so far, been formulated. For such 'laws' say no more and no less than this: The probability is almost 1 that a certain event will occur.

References

[1] Odd O. Aalen, Per Kragh Andersen, Ørnulf Borgan, Richard D. Gill, and Niels Keiding. History of applications of martingales in survival analysis. *Electronic Journal for History of Probability and Statistics*, 5(1), 2009. 4

[2] Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, editors. *Introduction to Imprecise Probabilities*. Wiley, 2014. 13

[3] Ole Barndorff-Nielsen, Preben Blæsild, and Geert Schou, editors. *Proceedings of Conference on Foundational Questions in Statistical Inference, Aarhus, May 7-12, 1973*. Institute of Mathematics, University of Aarhus, 1974. 23

[16] Wald surely did not mean to suggest that *every* event with probability near to one will happen. We may assume that he meant that easily describable events with probability near to one will happen, because he used a similar caveat in the 1930s to make Richard von Mises's axiomatization of probability mathematically rigorous [5, 34].

[4] J. H. Bennett, editor. *Statistical inference: Selected correspondence of R. A. Fisher*. Clarendon, Oxford, 1990. 12

[5] Laurent Bienvenu, Glenn Shafer, and Alexander Shen. On the history of martingales in the study of randomness. *Electronic Journal for History of Probability and Statistics (www.jehps.net)*, 5(1), 2009. 3, 22

[6] Leo Breiman. Optimal gambling systems for favorable games. In Jerzy Neyman, editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1 (Contributions to the Theory of Statistics), pages 65–78, Berkeley, CA, 1961. University of California Press. 5

[7] Marie-France Bru and Bernard Bru. *Les jeux de l'infini et du hasard*. Presses universitaires de Franche-Comté, Besançon, France, 2018. Two volumes. 18

[8] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, UK, 2006. 5

[9] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):943, 2015. 6

[10] Antoine Augustin Cournot. *Exposition de la théorie des chances et des probabilités*. Hachette, Paris, 1843. Reprinted in 1984 as Volume I (Bernard Bru, editor) of [11]. 1, 19

[11] Antoine Augustin Cournot. *Œuvres complètes*. Vrin, Paris, 1973–2010. The volumes are numbered I through XI, but VI and XI are double volumes. 23

[12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991. Second edition in 2006. 5

[13] David R. Cox. Some problems connected with statistical inference. *Annals of Mathematical Statistics*, 29:357–372, 1958. 15

[14] A. P. Dempster. The direct use of likelihood for significance testing. *Statistics and Computing*, 7(4):247–252, 1997. 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 [3] along with discussion by George Barnard and David Cox. 4, 10

[15] A. W. F. Edwards. *Likelihood. An account of the statistical concept of likelihood and its application to scientific inference*. Cambridge University Press, Cambridge, 1972. 4

[16] Ronald A. Fisher. On the 'Probable Error' of a coefficient of correlation deduced from a small sample. *Metron*, 1(4):3–32, 1921. 4

[17] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London (A)*, 222:309–368, 1922. 12

[18] Ronald A. Fisher. *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh, 1956. Subsequent editions appeared in 1959 and 1973. 4, 8

[19] Joseph Fourier. Mémoire sur les résultats moyens déduits d'un grand nombre d'observations. In Joseph Fourier, editor, *Recherches statistiques sur la ville de Paris et le département de la Seine*, pages ix–xxxi. Imprimerie royale, Paris, 1826. 1, 18

[20] D. A. S. Fraser, Nancy Reid, and Wei Lin. When should modes of inference disagree? Some simple but challenging examples. *Annals of Applied Statistics*, 12(2):750–770, 2018. 15

[21] Andrew Gelman and John Carlin. 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, 2017. 1, 9

[22] Gerd Gigerenzer. Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2):198–218, 2018. 1, 17

[23] Steven N. Goodman. Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, 130(12):995–1004, 1999. 17

[24] Trygve Haavelmo. The probability approach to econometrics. *Econometrica*, 12(Supplement):1–115, 1944. 1, 22

[25] Anders Hald. *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713 to 1935*. Springer, New York, 2007. 12

[26] David J. Hand. From evidence to understanding: A commentary on Fisher (1922) 'On the mathematical foundations of theoretical statistics'. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2039), 2015. 12

[27] Campbell R. Harvey. The scientific outlook in financial economics. *Journal of Finance*, 72(4):1399–1440, 2017. 6

[28] John L. Kelly Jr. A new interpretation of information rate. *Bell System Technical Journal*, 35(4):917–926, 1956. 5

[29] Solomon Kullback. *Information Theory and Statistics*. Wiley, New York, 1959. 4

[30] Tze Leung Lai. History of martingales in sequential analysis and time series. *Electronic Journal for History of Probability and Statistics*, 5(1), 2009. 4

[31] Erich L. Lehmann. *Fisher, Neyman, and the Creation of Classical Statistics*. Springer, New York, 2011. 17

[32] David G. Luenberger. *Investment Science*. Oxford University Press, New York, second edition, 2014. 5

[33] Andrei A. Markov. *Исчисление вероятностей*. Типография Императорской Академии Наук, St. Petersburg, 1900. The second edition, which appeared in 1908, was translated into German as *Wahrscheinlichkeitsrechnung*, Teubner, Leipzig, Germany, 1912. 2

[34] Per Martin-Löf. The literature on von Mises' Kollectivs revisited. *Theoria*, 35:12–37, 1969. 22

[35] Blakeley B. McShane and David Gal. Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, 112(519):885–895, 2017. 1

[36] Mary Morgan. *The History of Econometric Ideas*. Cambridge University Press, Cambridge, 1990. 21

[37] Jerzey Neyman. 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, 1937. 12, 15

[38] Jerzy Neyman. "Inductive behavior" as a basic concept of philosophy of science. *Review of the International Statistical Institute*, 25(1/3):7–22, 1957. 15

[39] Jerzy Neyman and Egon S. Pearson. On the use and interpretation of certain test criteria. *Biometrika*, 20A:175–240, 263–295, 1928. 7

[40] Glenn Shafer. Bayesian, fiducial, frequentist, 2017. Working Paper 50, `www.probabilityandfinance.com`. 16

[41] Glenn Shafer. Cournot in English, 2017. Working Paper 48, `www.probabilityandfinance.com`. 21

[42] Glenn Shafer and Vladimir Vovk. *Game-Theoretic Probability and Finance*. Wiley, Hoboken, New Jersey, 2019. 1, 13, 14, 15, 16

[43] Stephen M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, MA, 1986. 13

[44] Stephen M. Stigler. *Statistics on the Table: The Historyof Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA, 1999. 4

[45] Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939. 3

[46] Abraham Wald. *On the principles of statistical inference*. University of Notre Dame, 1942. Four lectures delivered at the University of Notre Dame, Februrary 1941. Printed by Edwards Brothers, Lithoprinters, Ann Arbor. 1, 21

[47] Ronald L. Wasserstein and Nicole A. Lazar. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016. 1

[48] Min-ge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: A review (with discussion). *International Statistical Review*, 81(1):3–77, 2013. 15

Acknowledgements

Many conversations over the past several years have inspired and influenced this paper. Most important, perhaps, were conversations about game-theoretic testing and meta-analysis with Peter Grünwald 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 Bock and their students at the University of Ghent and 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.

Vladimir Vovk has also provided useful feedback. My understanding of the idea of testing by trying to multiply the money one risks has developed over several decades in the course of my collaboration with Vovk.