The New England Journal of Statistics in Data Science Volume 2, 215–228 (2024)
DOI: https://doi.org/10.51387/23-NEJSDS55

Improving Data Analysis by Testing by Betting: Optional Continuation and Descriptive Statistics

Glenn SHAFER

Abstract

When testing a statistical hypothesis, is it legitimate to deliberate on the basis of initial data about whether and how to collect further data? Game-theoretic probability's *fundamental principle for testing by betting* says yes, provided that you are testing the hypothesis's predictions by betting and do not risk more capital than initially committed. Standard statistical theory uses *Cournot's principle*, which does not allow such *optional continuation*. Cournot's principle can be extended to allow optional continuation when testing is carried out by multiplying likelihood ratios, but the extension lacks the simplicity and generality of testing by betting.

Testing by betting can also help us with descriptive data analysis. To obtain a purely and honestly descriptive analysis using competing probability distributions, we have them bet against each other using the principle. The place of confidence intervals is then taken by sets of distributions that do relatively well in the competition. In the simplest implementation, these sets coincide with R. A. Fisher's likelihood ranges.

KEYWORDS AND PHRASES: Game-theoretic probability, Game-theoretic statistics, Optional continuation, Optional stopping, Cournot's principle, Fundamental principle of testing by betting, Ville's inequality, Descriptive statistics, Kelly betting, Likelihood, Probability forecasting, Convenience sample.

1. INTRODUCTION

Game-theoretic probability studies games in which announcements by Player I define rates at which Player II can bet on outcomes. The games can have one or many rounds. Player I's announcements can take various forms. One possibility is that Player I announces a probability distribution and Player II is authorized to buy any payoff for its expected value.

A probability distribution for a discrete-time stochastic process can be used as a strategy for Player I; it tells Player I what probability distribution for the nth outcome to announce after seeing the first $n-1$ outcomes. When Player I is required to follow a particular strategy of this form, strategies for Player II define martingales. Theorems about discrete-time stochastic processes become theorems in game theory [52, 54].

The betting intuition brought to the fore by game-theoretic probability has proven productive for statistical theory, sometimes helping mathematical statisticians find more efficient or more powerful methods [43]. More importantly, game-theoretic probability can help us go beyond today's standard mental framework for mathematical statistics, where we begin with a collection of probability distributions and test whether one of these distributions could have "generated" our data or try to decide which one did so. This article discusses two ways in which game-theoretic probability goes beyond the standard framework to help make data analysis more flexible and more honest.

Considered purely as mathematics, game-theoretic probability generalizes standard probability theory, because it does not assume that any of the players are required to follow a strategy specified in advance.[1] This allows us to authorize, in a simple and explicit way, deliberation on the basis of initial data about (1) whether and how to collect further data and (2) how to use it in statistical testing. We already see such deliberation in practice, but squaring it with standard probability theory is not so simple and limits the tests authorized to those used in the game-theoretic picture. I discuss this point in detail in Section 2.

In Section 3, I consider a second way game-theoretic probability can help us improve data analysis. Ever since statistical testing began to be widely used in the early 19th century, its greatest abuse has been unjustified and often clearly erroneous assumptions of randomness. Sometimes the assumption is that successive observations are chosen randomly from a population. Sometimes the assumption is that the deviations of successive observations from a model are random – i.e., independent of each other and of the ex-

arXiv: 2308.14959

[1] Game-theoretic probability also allows Player I to make more limited betting offers, such as those considered in the theory of imprecise probabilities [2, 56] or more comprehensive betting offers, such as those considered by [57]. I do not study these aspects of game-theoretic probability's flexibility in this paper.

planatory variables in the model. In some cases, practitioners acknowledge that the assumptions are unjustified but claim that their analyses are nevertheless interesting as descriptions of data. Because the statistical language being used is at least implicitly causal, this usually sounds like double talk. The betting picture can help here by providing a language that is clearly descriptive rather than causal. We rank the probability distributions in a model by having them bet against each other's predictions. This produces a set of distributions that did best in the competition. This set is not a confidence set. There is no suggestion that we are confident about anything, and we are certainly not confident that one of the distributions in the set "generated" the data. The set is simply a set of distributions that predicted the data better than the other distributions in the model. The betting language provides an intuitive scale for this predictive success.

These two improvements in data analysis are not so much improvements in what we do as improvements in how we understand and explain what we do. In practice, researchers already deliberate about how to continue experimentation. Some already use the descriptive sets we obtain in Section 3; they are called "likelihood ranges". Aside from setting aside some types of testing that cannot be legitimately authorized when we deliberate about continuation, the proposed improvements concern how we can communicate data analyses more clearly and more honestly to ourselves and to others. But it is also notable and important that the betting language allows us to generalize seamlessly to the case where predictions being tested or compared are not necessarily produced by probability models. They may instead be produced by algorithms of a very different kind — perhaps neural nets, perhaps physical models like those used in weather forecasting.

2. STATISTICAL TESTING WITH OPTIONAL CONTINUATION

Statistical testing requires a mathematical theory of probability together with a principle that specifies how probabilities can be discredited by observations.

- The principle used to make traditional probability theory into a theory of statistical testing is sometimes called *Cournot's principle*.[2] This principle authorizes a statistician to select an event to which a probability distribution assigns small probability and to regard its happening as evidence against the distribution.
- To make game-theoretic probability into a theory of statistical testing, we can use a principle that I have called *the fundamental principle for testing by betting*.[3] This principle, which is related to but distinct from Cournot's principle, authorizes a statistician to interpret success in betting against a probability distribution as evidence against the distribution.

Do these principles authorize optional continuation?

As the term is used here, *optional continuation* refers to the practice of deliberating, after seeing some initial data, about whether and how to continue collecting and analyzing data. Such continuation may involve observations or experiments not contemplated at the outset. It is distinguished from optional stopping, which refers in established usage only to the possibility of adopting at the outset a plan specifying circumstances under which we will curtail a fully planned sequential experiment or observational study.

The fundamental principle for testing by betting asserts the validity of optional continuation for the type of testing it considers. Cournot's principle, in its classical formulation, does not. It can be extended to assert the validity of optional continuation when testing is carried out by multiplying likelihood ratios, but as I will explain, the extension lacks the simplicity and generality of the fundamental principle for testing by betting.

2.1 Optional Continuation in Practice and Theory

Optional continuation has long been part of statistical practice. It is implicit, for example, in the idea of meta-analysis. But it has proven difficult to bring it under the purview of statistical theory.

The term "optional continuation" with the meaning used here first appeared in print in Allard Hendriksen's master's thesis at the University of Leiden, written under the supervision of Peter Grünwald [33]. Hendriksen wrote on page 3 of the thesis,

> "Optional continuation" is the practice of combining evidence of studies that were done because of promising results of previous research on the same subject.

The term has subsequently been used in other work by Günwald's machine-learning research group at CWI in Amsterdam [32, 31]. But as of June 13, 2023, it had not yet appeared in any of the 34 statistics journals in JSTOR.

The older term "optional stopping" was introduced by the Duke mathematician Joseph Albert Greenwood [29]. Greenwood sought empirical adjustments to account for the way Joseph Rhine's laboratory was conducting and analyzing its experiments on extra-sensory perception. Rhine stopped ex-

[2] See [53] and [46, 51] for the history of Cournot's principle. The principle is sometimes ridiculed by philosophers; see for example [35, p. 49] and [14, pp. 66–67]. But it has been articulated in one way or another by a panoply of mathematicians and statisticians, including Jacob Bernoulli, Antoine-Augustin Cournot himself, Émile Borel, Richard von Mises, Andrei Kolmgorov, Abraham Wald, Joseph L. Doob, William Feller, Harold Jeffreys, Charles Stein, and Philip Dawid.

[3] Vovk and I have used various other names for this principle. In 2001, we called it *the fundamental interpretative hypothesis of probability* [52, pp. 5, 14, 62]. In 2019, we called it *the game-theoretic version of Cournot's principle* [54, pp. 226–227].

perimenting with each subject when a success rate thought to be statistically significant was achieved, then combined the z-scores achieved by successive subjects.

Greenwood's problem was brought to wider attention in mathematical statistics by William Feller's critiques of the ESP work [20, pp. 272, 286–292], [19, pp. 140, 190, 197]. In subsequent work in probability, "optional stopping" has referred to stopping rules that can be adopted in advance without annulling a desired property of a stochastic process, usually the property of being a martingale [16].

In his book on sequential analysis [59], Abraham Wald considered only "sequential sampling plans" chosen in advance. While allowing early stopping when there was enough evidence to make a decision, these plans specified whether or not to stop and how to continue if stopping was not mandated, all as a function of outcomes so far. In a review of the book, George Barnard wrote that sequential analysis marked "the entry of statistical considerations into the very process of experimentation itself" [3]. We know that the process of experimentation often involves not only plans adopted in advance but also opportunistic changes in plans, based on new insights and unexpected information.

Barnard seems not to have followed up on his insight concerning the role of statistics in the process of experimentation; he did not discuss it, for example in his major article on statistical inference [4]. But in a subsequent article entitled "Sequential experimentation", R. A. Fisher wrote about the need for sequential deliberation in these terms [21, p. 183]:

> The present use of the term sequential is intended to be of a broader import than the formal use of the word as associated with the systematic procedure known as sequential analysis. The experimenter does not regard his material as wholly passive but instead looks to what may be learnt from it with a view to the improvement and extension of the enquiry. This willingness to learn from it how to proceed is the essential quality of sequential procedures. Wald introduced the sequential test, but the sequential idea is much older. For example, what is the policy of a research unit? It is that in time we may learn to do better and follow up our more promising results. The essence of sequential experimentation is a series of experiments each of which depends on what has gone before. For example, in a sample survey scheme, as explained by Yates, a pilot survey is intended to supply a basis for efficiently planning the subsequent stages of a survey. . . .

Until the recent work on optional continuation, this insight about statistical practice has remained outside the ambit of statistical theory.

2.2 A Betting Game with Optional Continuation

The simplest game used in game-theoretic probability has three players: Forecaster makes probability predictions, Skeptic bets against them, and Reality announces the outcomes. The game is a perfect-information game, meaning that the players move in turn and see each other's moves. We can vary the rules of the game, but we need not impose any further condition on what information any player might have or acquire in the course of the game, or on how the players might collaborate. Forecaster and Skeptic might be the same person. Forecaster and Reality might be the same person.

If Forecaster keeps forecasting, Skeptic can keep betting. Forecaster need not follow a plan or strategy about what to forecast next or how to forecast it.[4] Even if Forecaster follows a strategy known to Skeptic, Skeptic need not have a plan or strategy for when or how to bet against the forecasts. Thus optional continuation is built into the game, for both Forecaster and Skeptic. Skeptic can decide whether and how to continue selecting from Forecaster's betting offers, but Forecaster can decide what experiments or observations to make and what forecasts (perhaps probabilities) to give for them.

Vovk and I have used the example of quantum mechanics to illustrate game-theoretic probability's capacity for optional continuation; see [52, pp. 189–191] and [54, pp. 215–217]. In this example, we split Forecaster into two players, Observer and Quantum Mechanics. Observer selects the experiment, and Quantum Mechanics makes the probability forecast. Formally, the game continues indefinitely, and both Observer and Skeptic can effectively stop it by making null moves.

Although optional continuation is built into the game, we need this principle to use the game in statistical testing:

Principle 1 (Fundamental principle for testing by betting)**.** *Successive bets against a forecaster that begin with unit capital and never risk more discredit the forecaster to the extent that the final capital is large.*[5]

In one sense, this says it all. But some elaboration may be useful:

1. The principle is *fundamental*, not the consequence of some more extensive philosophy or methodology. We do not begin by saying that the forecaster's probabilities are or should be objective, subjective, personal, "frequentist", or whatever. We are testing the forecaster qua forecaster, and so we are testing his forecasts qua forecasts; the question is only whether they are good forecasts, relative to the knowledge and skill of whoever is doing the testing.
2. The forecaster may give a probability for a single event A, a probability distribution for an outcome X, or something less than a probability or a probability distribution:
   - If the forecaster gives a probability, you may bet on either side at the corresponding odds.

[4] To see how probability's limit theorems can be generalized to accommodate Forecaster's freedom, see [54, §7.5].

[5] I first formulated the principle in these words in my SIPTA lectures [49].

- If the forecaster gives a probability distribution for X, you may buy or sell any payoff $S(X)$ for its expected value.
- If the forecaster gives only an estimate E of X, you may buy or sell X for E.
- If the forecaster repeatedly gives a new probability for A or new estimate for X, say daily, you may buy or sell tomorrow's price for today's price.
- If the forecaster gives upper and lower previsions, you may buy at the upper or sell at the lower.

3. You *begin with unit capital* only for mathematical convenience. The discredit is measured by the ratio (final capital)/(initial capital), which I call the *betting score*.[6]
4. If you make several bets against the same forecaster (or the same theory or closely related theories), each starting with its own capital, then you are not allowed to report only the cases where you discredited the forecaster. Instead, you must report the overall result, the sum of your final capital over all the bets divided by the sum of your initial capital over all the bets.
5. When betting against successive forecasts, each bet uses only the cumulative capital remaining after the previous bet. You may not borrow or otherwise raise more capital in order to continue betting. This is what *never risk more* than the initial capital means.
6. When you stop, you must compare your initial capital with your *final* capital. You cannot claim to have discredited the forecaster because you had reached a higher level of capital in the interim. You do not have the money if you kept betting and lost it.[7]

I have stated the fundamental principle for testing by betting in 26 words, then taken a page to explain it. Is the principle simple? In any case, it is coherent and teachable. In contexts where the forecasts are only single probabilities or estimates, the principle can be taught even to those who have never studied mathematical probability. Moreover, the principle builds on ideas about betting that most people acquire before ever studying mathematical probability. Too many predictions contradicted by experience discredit the person making them. If you lose too much money betting on something, you are not much of an expert about it. Etc.

2.3 Cournot's Principle in Classical Form

What principles must we add to traditional probability theory to allow optional continuation?

Before addressing this question, I discuss a more basic question: How are we authorized to discredit a probability distribution P using observations? As I have already mentioned, the classical answer is Cournot's principle: we select an event E that has small probability $P(E)$ (call E our *test event*). The probability distribution P is discredited if E happens; we prefer to believe that the probabilities are incorrect rather than think that this improbable event happened.

Principle 2 (Cournot's principle)**.** *If we specify an event E in advance, and E happens, then we may take α, the probability of E, as a measure of evidence against P. The magnitude of discredit is measured by how small α and thus how large $1/\alpha$ is.*

We may call $1/\alpha$ our *test score*:

$$\text{test score} = \begin{cases} 1/\alpha & \text{if } E \text{ happens} \\ 0 & \text{if } E \text{ does not happen.} \end{cases} \tag{2.1}$$

Cournot's principle can be considered a special case of the fundamental principle for testing, because $1/\alpha$ is the capital that would result from E's happening if you bet unit capital on E. The test score (2.1) is also a betting score.

Although Cournot's principle has long been fundamental to statistical theory, current philosophical fashion has made it difficult to teach. A frequent objection is that some event of small probability always happens. When we hear this objection, we emphasize "specified in advance", which requires less emphasis when testing by betting, because a bet must be made in advance.

In some cases, we say "simple to describe" instead of or in addition to "specified in advance". Simplicity is also implicit to some extent when testing by betting, because a bet cannot be made and implemented unless the event is relatively simple.

2.4 Extending Cournot's Principle to Test Variables

This extension of Cournot's principle does not require us to specify in an advance a goal $1/\alpha$ for the strength of the evidence.

Suppose S is a nonnegative random variable, chosen in advance and so not too hard to describe, with expected value $E_P(S) = 1$ (call S our *test variable*). Our next principle says that a realized value s of S discredits P to the extent that s is much larger than 1.

Principle 3 (Authorization to test with a test variable)**.** *If we specify a test variable S in advance, then we may take s, the observed value of S, as a measure of evidence against P. We interpret s (our test score) on the same scale as we use in Cournot's principle. In other words, when $s = 1/\alpha$, it has the same force against P as the happening of a pre-specified event E when $P(E) = \alpha$.*

Cournot's principle is the special case of Principle 3 where S is given by (2.1). Principle 3 adds to Cournot's principle

[6] The alternative name *e-value*, designed to resemble *p-value* and avoid reference to betting, has recently become popular [58, 43].

[7] The anonymous 13th-century author who left us with the earliest surviving calculation of the chances for a throw of three dice warned us [34, p. 172]: "Addeque, quod lusor se continuare lucrando nescit, perdendo nescit dimittere ludum." Not knowing how to maintain his luck when winning, the gambler does not know how to quit when losing.

the possibility of a more graduated report on the strength of the evidence against P. The test scores it provides can be compared with p-values, which also provide a graduated report. A p-value p discredits P to the extent that it is small or $1/p$ is large. But a given value of $1/p$ carries less force than the same value s of a test score from a test variable. See §2.8.

It might seem that the greater flexibility offered by a test variable S comes at a price. When s is the realized value, the events $\{S = s\}$ and $\{S \leq s\}$ happen, and Markov's inequality tells us that our score $1/P(E)$ would have been at least as great, often greater, had we chosen one of these events as our test event E. But of course we could not have made these choices, because we did not know s in advance.

Like the classical form of Cournot's principle, Principle 3 can be considered a special case of the fundamental principle for testing by betting. The observed value s of the test variable S is the capital that would result from buying S for its expected value. The test variable S is our bet, and s is our betting score.

2.5 Extending Cournot's Principle to Test Martingales

Now suppose we want to test a probability distribution P for a stochastic process $X := X_1, X_2, \ldots$, and we observe the X_t successively. We use a *test martingale*, a nonnegative martingale $S_1, S_2, \ldots$ with $E_P(S_1) = 1$, again chosen in advance and hence relatively simple. The value s_t of S_t may become known to us only when we have observed $X_1, \ldots, X_t$. To interpret s_t, we adopt this principle:

Principle 4 (Authorization to test with a test martingale)**.** *If we specify a test martingale $S_1, S_2, \ldots$ in advance, then at all times t we may take s_t, the observed value of S_t, as the current measure of evidence against P. We interpret each s_t (each test score) on the same scale as we use in Principles 2 and 3. In other words, when $s_t = 1/\alpha$, it has the same force against P as the happening of a pre-specified event E with $P(E) = \alpha$.*

Principle 3 is the special case of Principle 4 where all the S_t for $t \geq 2$ are equal to the constant 1.

Like Cournot's principle and our previous extensions of it, Principle 4 can be considered a special case of the fundamental principle of testing by betting. For each t, s_t is our capital after observing X_t if we first buy S_1 and then, for each time $u > 1$, invest all our capital after observing X_u in S_{u+1}. But we are still testing a mathematical object, a probability distribution P. Forecaster is following a fixed strategy, which tells him to use P's successive conditional probabilities as forecasts, and Skeptic's strategy (the test martingale) is also specified in advance. Neither Forecaster nor Skeptic have any discretion after the process of begins. So Principle 4 is not a principle of optional continuation in the sense of this article. It allows optional stopping only in Doob's sense; before the process begins, Skeptic can modify a proposed test martingale by adopting a plan for stopping.

Although the statement of Principle 4 does not mention betting, I do not recall seeing the principle explained without a betting story. I do not know any other way to motivate it.

2.6 Improvised Testing (Optional Continuation for Skeptic)

Principle 4 authorizes the statistician to use a test martingale specified in advance. Improvisation is not yet authorized. For this, we need some further principle. As with Principle 4, we are testing a probability distribution P for a stochastic process $X := X_1, X_2, \ldots$, and we observe the X_t successively. When $x_1, \ldots, x_{t-1}$ are possible values of $X_1, \ldots, X_{t-1}$, we call a nonnegative variable $S(X_t)$ a *round-t test variable given $x_1, \ldots, x_{t-1}$* if $E_P(S(X_t)|x_1, \ldots, x_{t-1}) = 1$; when $t = 1$, this reduces to $E_P(S(X_1)) = 1$. We can formulate a principle for improvisation in testing as follows:

Principle 5 (Authorization to improvise when testing)**.** *Suppose we set $s_0 = 1$, specify a round-1 test variable, say $S_1(X_1)$. Then, beginning with $t = 1$,*

1. *we observe X_t's value x_t,*
2. *we set $s_t := s_{t-1}S_t(x_t)$, and*
3. *we specify a round-$(t+1)$ test variable given $x_1, \ldots, x_t$, say $S_{t+1}(X_{t+1})$.*

Suppose we continue so long as we still want to continue and stop whenever we please (after step 2 for some t). Then at all times t until after we stop, we may take s_t as the current measure of evidence against P. We interpret s_t on the same scale as we use in Principles 2, 3, and 4.

Principle 5 generalizes Principle 4, and like Principle 4, it can be considered a special case of the fundamental principle for testing by betting. Skeptic is now a free player, not constrained to follow a strategy specified in advance.

2.7 Improvised Probabilities (Optional Continuation for Forecaster)

Principle 5 authorizes a statistician testing a probability distribution to improvise. But this still does not bring us to R. A. Fisher's vision, where a statistician helps construct over time not only a test but also the probabilities being tested. In this vision, the statistician brainstorms with other scientists to design an experiment with outcome X_1, to which they assign probabilities based on some theory they want to test, and after observing $X_1 = x_1$, they brainstorm again about what they have learned and design a possibly unanticipated experiment with outcome X_2, and so on.

It is tempting to try to square traditional probability with Fisher's vision by imagining that this collaboration defines a probability distribution P progressively. on the first step we define a probability distribution P_1 for X_1. On the second,

we define a probability distribution P_2 for X_2, and so on. We are tempted to say that we are testing the product $P_1 \times \cdots \times P_k$, where k is where the research team stops. But the statistician did not set out to test $P_1 \times \cdots \times P_k$. She and her colleagues waited to design the second experiment and its X_2 and P_2 until they had seen x_1. Had x_1 come out differently, their subsequent brainstorming might have produced a different X_2 and P_2, and so on. If there is a single comprehensive probability distribution being tested, it is not $P_1 \times \cdots \times P_k$; instead it must give conditional probabilities for each X_t given all the different ways the previous xs might have come out and all the different ways the research team's information and thinking might evolve while the previous experiments were being performed and analyzed.

Several decades ago Philip Dawid [12, 13] bravely argued that these dependencies should not matter—that we can design significance tests, confidence intervals, and Bayesian procedures that are unaffected by probabilities, somehow true or somehow invented, involving the might-have-beens. As these might-have-beens do not matter, we can just pretend that we have the desired independence. This is Dawid's *prequential* model. Although some statisticians (including myself) found it appealing, others found it confusing. What are we really testing? Are we testing a huge and not fully specified probability distribution P whose unspecified probabilities include probabilities for actions of the research team doing the testing? In my experience, some mathematicians say yes, but the formulation has a complicated and paradoxical air that hardly lends itself to communication with scientists and their public.

Leaving aside the problem of communication, can we formulate a principle that authorizes us to use Dawid's insight to construct test scores? Here's a try.

Principle 6 (A prequential testing principle)**.** *Suppose we set $s_0 = 1$, construct an experiment that will produce a variable X_1, select a probability distribution P_1 for X_1, and select a test variable S_1 for P_1. Then, beginning with $t = 1$,*

1. *we observe X_t's value x_t,*
2. *we set $s_t := s_{t-1}S_t(x_t)$, and*
3. *we construct an experiment (perhaps newly conceived) that will yield a variable X_{t+1}, a probability distribution P_{t+1} for X_{t+1}, and a test variable S_{t+1} for P_{t+1}.*

Suppose we continue so long as we still want to continue and stop whenever we please (after step 2 for some t). Then at all times t until after we stop, we may take s_t as the current measure of evidence against the P_t we have constructed so far all being valid. We may interpret s_t on the same scale as we use in Principles 2, 3, 4, and 5.

Principle 5 is a special case of Principle 6. And Principle 6, like our preceding extensions of Cournot's principle, can be considered a special case of the fundamental principle for testing by betting. Now both Forecaster and Skeptic are free agents, not constrained to follow any strategy specified in advance.

The principle's consistency with testing in the game-theoretic framework is not surprising, as that framework was partly inspired by Dawid's prequential model.

2.8 The Role of Ville's Inequality

Ville's inequality says that if $S_1, S_2, \ldots$ is a test martingale, then

$$P\left(\sup_{t\geq 1} S_t \geq \frac{1}{\alpha}\right) \leq \alpha$$

for every $\alpha > 0$. Some people (including myself) have sometimes said that Ville's inequality authorizes optional continuation. This is a careless formulation. First because a theorem is never more than mathematics; it cannot authorize anything. Second because the principle Ville's inequality suggests is not an optional continuation principle in the sense developed in this article. It begins with the choice of a test martingale and so does not help us understand optional continuation for Skeptic or optional continuation for Forecaster.

Recall that a random variable W satisfying $P(W \leq \alpha) \leq \alpha$ for all $\alpha > 0$ is called a *p-variable*, and that a realized value of a p-variable is called a *p-value*.[8] Ville's inequality tells us that $1/\sup_{t\geq 1} S_t$ is a p-variable, and so $1/\sup_{t\geq 1} s_t$ is a p-value. Well, almost. It is at least implicit in the notion of a p-value, as statisticians understand and use the term, that we have observed it and know we have observed it. We do not expect this for $1/\sup_{t\geq 1} s_t$. But we do observe upper bounds. At time t, we have observed the upper bound $1/\sup_{1\leq i\leq t} s_i$, and an upper bound on a p-value is a p-value. So most statisticians who use p-values would probably accept this principle:

Principle 7 (The dynamic p-value principle)**.** *At any time t as we observe a sequence $S_1, S_2, \ldots$ that is a martingale under P, we may interpret $1/\sup_{1\leq i\leq t} s_i$ as evidence against P just as statisticians usually interpret a p-value.*

Principle 7 can be called an optional stopping principle. It is not an optional continuation principle in the sense of this article, because it does not authorize us to change the test martingale or later experiments and the probabilities for their Xs.

Because the $1/\sup_{1\leq i\leq u} s_i$ for $u < t$ are no smaller than $1/\sup_{1\leq i\leq t} s_i$, these earlier $1/\sup_{1\leq i\leq u} s_i$ still have the force of a p-value against P when we give $1/\sup_{1\leq i\leq t} s_i$ the force of a p-value against P. This observation is related to the concept of a confidence sequence, which goes back to [11].

As an optional stopping principle for martingales, Principle 7 can be compared with Principle 4. Neither is stronger than the other. Principle 4 authorizes us to use s_t as a measure of our evidence against P and to continue doing so if

[8] See [55, p. 88] for an explanation of how this definition of *p-value* is equivalent to the traditional definition in terms of test statistics. In many branches of theoretical and applied statistics, the distinction between *p-variable* and *p-value* is ignored; both are called p-values.

we stop. But it does not allow us to continue using s_t if we do not stop and hence does not authorize us to use the sometimes larger $\sup_{1\le i\le t} s_i$ [55]. But it gives $1/s_t$ the force of a fixed significance level, which is greater than the force of a p-value. When we observe a p-value p, we are observing the happening of the event $W \le p$, which has probability p or less. But we did not specify this event in advance; we only specified the p-variable W.

Ville's inequality and Principle 7 have generalizations in game-theoretic statistics, where they use game-theoretic definitions of upper and lower probability and expected value [54, Exercise 2.10].

2.9 Conclusion

We have discussed how Cournot's principle, the traditional principle for testing a probability distribution, can be extended so that it fully accommodates optional continuation and yet does not explicitly use game-theoretic probability or ideas about betting. Is this extension worth the trouble?

The clear message of the exercise is that the fundamental principle of testing by betting, coupled with game-theoretic probability, provides a theoretical basis for optional continuation that is simpler, clearer, and more general. Readers will judge for themselves, but I submit that Principle 6 is overly complex, ill-motivated, and impossible to teach without reference to betting. It remains, moreover, less general than the fundamental principle of testing by betting, because it requires Forecaster's moves on each round of a forecasting game to be a probability distribution.

Our exercise has also illustrated the new clarity brought to statistical theory by game-theoretic probability's distinction between Forecaster and Skeptic. This distinction has helped us see the complexity of the notion of optional continuation. Optional continuation for Forecaster is a step further than optional continuation for Skeptic.

3. GAME-THEORETIC DESCRIPTIVE DATA ANALYSIS

Researchers often construct statistical models that cannot be taken seriously as anything more than descriptions of their *study populations* — the populations for which they have data. Unfortunately, our methodology and terminology for constructing such models (estimation, significance tests, confidence intervals, credible regions, etc.) can only be understood in terms of inferences about larger populations or observations not yet made.

A study population is often merely a *convenience sample* (examples we managed to find). Sometimes it is an *entire population* (perhaps the eight highly industrialized nations, or the five hundred corporations in the S&P 500 index). Describing such populations means summarizing — identifying general features rather than details.

When a study population is merely a collection of numbers, we may be able to summarize it by giving a few descriptive statistics, such as the average $\overline{y}$ and the standard deviation s. But confusion arises as soon as we ask about the precision of these statistics. If $\overline{y} = 5.346$, then is 5 just as good a description? What about 6 or 4? Maybe. If the numbers are very spread out, then saying that 0 is in the middle might be just as good as saying that 5.346 is in the middle. Our usual response to this difficulty is to replace $\overline{y}$ with a confidence interval, such as $\overline{y}\pm 1.96s/\sqrt{n}$, but then we are pretending to make an inference to a theoretical mean or a larger population.

The first thesis of this section is that when our goal is merely to describe a study population, we should talk about prediction, not about inference. The average of a collection of numbers can be used to predict each number, and asking whether alternatives to the average predict the numbers in the collection as well or nearly as well does not sidetrack us into thinking about inference to some larger population. The predictions considered here are probability distributions. Following [54], I will call them *probability forecasts* or simply *forecasts*.

The second thesis is that betting can give us a scale for comparing probability forecasts that has intuitive meaning without recourse to inferential ideas. The key is to have the different probability distributions in a model (or different algorithms that produce probability forecasts) bet against each other. The factor by which one distribution or algorithm multiplies the money it risks betting against another is a measure of how much better it did as a forecaster, a measure that has an intuitive meaning even for those not trained in probability theory.

3.1 Theory

As explained in Section 2, game-theoretic probability uses probability distributions as forecasting strategies. It recasts the notion of independent observations as a property a forecasting strategy might have: the strategy always makes the same forecast or uses the same forecasting rule. The question whether given observations are random with respect to a probability distribution is replaced by the question whether such a forecasting strategy withstands bets against its forecasts.

In addition to serving as a forecasting strategy, a probability distribution can also serve as a strategy for betting against a probability forecaster. This was explained by John L. Kelly Jr. [37], and the strategy has been called *Kelly betting*; see also [54, Ch. 10] and [50]. In the simplest case, Kelly betting produces a likelihood ratio as the payoff of a gamble.

The duality in the way probability distributions can be used — as forecasting strategies and as strategies for betting against forecasts, is key to the simple idea I am proposing here: Choose a statistical model (i.e., a collection of probability distributions), have them bet against each other's

forecasts for the study population, and take the distribution or distributions that do best in the competition as our description of the study population. The result is purely descriptive; it does not suppose that the winners will forecast well in other data, and it does not even make any claims about the model we have chosen being best for the study population. The choice may be purely conventional.

3.1.1 From Probability Forecasts to Description

Any method of description requires relatively arbitrary choices and conventions. Here we begin by choosing variables that have values for each individual in our study population: a variable Y (the *target* variable) and variables $X_1, \ldots, X_K$ (the *forecasting* variables). Inferential statistical theory has accustomed us to think of these choices as causal modeling, but when description is our goal, Y and the X_k are simply what we want to describe. For whatever reason, we want to know how they are related in the study population.

Next we choose a family of algorithms $(P_\theta)_{\theta\in\Theta}$, where each algorithm P_θ uses the X_k to give probability distributions for Y. The P_θ are our forecasters, and the probability distributions they give are our forecasts.

We will want to limit the complexity of the P_θ. In inferential statistics, simplicity is said to be a virtue because complex forecasters are likely to overfit — i.e., fail to generalize beyond the study populations. When are goal is description rather than inference, either because we are uninterested in other populations or because we cannot make assumptions that would justify inference to other populations, a more immediate virtue of simplicity is salient. Description is description only when it is simple enough to be understood.

Being familiar and conventional is another virtue of a descriptive forecasting family. Description requires convention, and it can only be communicated to those who know the convention.

Once the variables and the forecasting family are chosen, we can evaluate the θ according to their relative forecasting success within the study population. Let Θ_{D} be the subset of Θ consisting of θ whose forecasts perform reasonably well in our judgment; we may call it the *description range*. Its elements are the *descriptive forecasters of* Y. Often we will be most interested in some particular aspect of the descriptive forecasters. When $\theta = (\mu, \sigma^2)$, for example, we might be interested in μ. In other cases, we might be interested in the difference $Y_{\mathrm{A}} - Y_{\mathrm{B}}$ when values of the X_k for two individuals A and B are given. In general, the range of $h(\theta)$ when θ ranges over Θ_{D} is our *description range* for $h(\theta)$.

3.1.2 The Betting Competition

For each ordered pair (θ_1, θ_2) of elements of Θ, we pit θ_2 bet against θ_1. Forecaster predicts Y with the distribution P_{θ_1}; Skeptic predicts Y with the distribution P_{θ_2}. Skeptic has unit capital. Forecaster offers to sell Skeptic any nonnegative payoff $S(Y)$ for $E_{\theta_1}(S)$, his expected value for S. In deciding what nonnegative payoff S to buy, Skeptic can use his distribution P_{θ_2} in different ways, but when he uses Kelly betting he buys the payoff S given by

$$S(Y) := P_{\theta_2}(Y)/P_{\theta_1}(Y) \tag{3.1}$$

at the price Forecaster requires:

$$E_{\theta_1}\left(\frac{P_{\theta_2}(Y)}{P_{\theta_1}(Y)}\right), \text{ which is equal to 1.}$$

This bet turns Skeptic's initial unit capital into $S(y) = P_{\theta_2}(y)/P_{\theta_1}(y)$. This ratio, the betting score, provides a measure of θ_2's forecasting success relative to θ_1.

For our descriptive purposes, Kelly betting can be considered one more convention. It does, however, have well known optimization properties for Skeptic when Skeptic is confident in Q as a forecast.[9] An important alternative to Kelly betting is fractional Kelly betting, which risks only a fraction of one's capital. Being more cautious, this penalizes Skeptic less when $P_{\theta_2}(y)/P_{\theta_1}(y)$ is low. This makes the competition between forecasters less sensitive to individuals that are unusual with respect to the whole study population.

The ratio $P_{\theta_2}(y)/P_{\theta_1}(y)$ is familiar to statisticians under the name *likelihood ratio*. We will usually chose our forecasting family so that a unique maximum of $P_\theta(y)$ always exists. The value of θ that achieves the maximum, say $\hat\theta$, is the *maximum-likelihood estimate* when y is observed, and the likelihood ratio

$$L(\theta) = \frac{P_\theta(y)}{P_{\hat\theta}(y)}$$

is a number between zero and one that measures θ's forecasting performance.

Using cutoffs suggested by R. A. Fisher in 1956,[10] we may classify the performance of the θ according to their value of $L(\theta)$ as follows:

Very good	$L(\theta) \geq 1/2$
Good	$1/2 > L(\theta) \geq 1/5$
Satisfactory	$1/5 > L(\theta) \geq 1/15$
Unsatisfactory	$1/15 > L(\theta)$

(3.2)

The names in (3.2) are my suggestions; Fisher did not provide names for the intervals. Whatever names and cutoffs we choose will be arbitrary conventions, but no more arbitrary than the terminology and the cutoffs 5% and 1% used for statistical inference. If equally accepted as conventions, they can be equally serviceable. The meaning in terms of betting will be readily understood by the public, including those not trained in mathematical probability.

For Fisher, the intervals were inferences. A number of other authors, beginning with A. W. F. Edwards [17] and Richard Royall [45], have adopted Fisher's proposal to use

[9] See [8]. For more on Kelly betting, see [18, Chapter 10] and [60]. For other roles Kelly betting can play in statistical theory, see [50] and [58].

[10] See [22, p. 71] or page 75 of the posthumous third edition (1973).

$L(\theta)$ for inference. These authors have developed methods for computing, tabulating, and displaying $L(\theta)$, intervals of its values, and other summaries for a large variety of models. I will not attempt to review this very extensive work, but it is obviously an asset for the descriptive proposal I am making here.

Fisher called $L(\theta)$ the *likelihood* of θ. This name has endured for a century, and no mathematical statistician will be able to put it out of mind while reading the rest of this paper. Yet I will avoid it as much as possible, because its inferential connotation cannot be circumvented. On the other hand, I am using the familiar notation: $\hat{\theta}$, $L(\theta)$, and also $l(\theta)$ for $\ln(L(\theta))$.

3.2 Examples

I offer three very simple examples. The first is purely formal and about as simple as possible: the forecast is a single probability, always the same probability. The second involves a classic convenience sample. The third, a fictional instantiation of a problem of current interest, involves an entire population.

3.2.1 Forecasting with a Single Probability

Suppose we observe successive trials of an event, and each algorithm in our forecasting family has a fixed probability that it uses each time as its forecast. Formally, $\Theta = [0, 1]$, and Forecaster θ always gives θ as its forecast.

If we observe 100 trials, and the event happens 70 times, then $\hat{\theta} = 0.7$, and

$$L(\theta) := \left(\frac{\theta}{0.7}\right)^{70} \left(\frac{1-\theta}{0.3}\right)^{30}.$$

Our scheme for rating the forecasters yields these approximate description ranges:

Very good	$0.64 < \theta < 0.76$
Good	$0.61 < \theta < 0.78$
Satisfactory	$0.59 < \theta < 0.80$

The forecaster $\theta = 1/2$ may have been of particular interest, and we may want to emphasize that its performance was unsatisfactory.

Not surprisingly, Fisher's categories are roughly consistent with inferential practice. The standard error of the maximum-likelihood estimate 0.7 is 0.046, suggesting a 95% confidence interval of $(0.61, 0.79)$. Like this confidence interval, our description ranges do not contain 1/2. But unlike the confidence interval, the description ranges merely describe; they merely tell us which constant forecasts perform relatively well in the data. This does not involve attribution of independence in any sense to the trials themselves.

3.2.2 Fourier's Masculine Generation

The calculation of error probabilities from statistical data was first made practical by Laplace's central limit theorem, and the calculation was explained to statisticians by Joseph Fourier (1768–1830). Fourier had been an impassioned participant in the French revolution and an administrator under Napoleon. After the royalists regained power, a prominent royalist who had been Fourier's student, Chabrol de Volvic, rescued him from impoverishment with an appointment to the Paris statistics bureau. This assignment left him time to perfect the theory of heat diffusion for which he is best known, but as part of his work at the statistics bureau, he published two marvelously clear essays on the use of probability in statistics, in 1826 and 1829. According to Bernard Bru, Marie-France Bru, and Olivier Bienaymé, these were the only works on mathematical probability read by statisticians in the early 19th century.[11]

To illustrate Laplace's asymptotic theory, Fourier studied data on births and marriages gleaned from 18th-century parish and governmental records in Paris. He was particularly interested in the length of a masculine generation — the average time, for fathers of sons, from the father's birth to the birth of his first son. On the basis of 505 cases, he estimated this average time to be 33.31 years. In his bureau's report for 1829 [24, Table 64, p. 143ff], he gave the bounds on the estimate's error shown in Table 1.

Laplace's theory is applicable, of course, only if the 505 cases are a random sample, and they are not. They are a convenience sample, consisting of 18th-century fathers of sons in Paris for which the needed parish records could be found. This convenience sample is of interest, but Fourier's inferential analysis of it is unjustified. A descriptive analysis is needed.

For description, we do not need Fourier's implicit assumption that the 505 cases were a random sample. We need a conventional forecasting family. Let us use the most conventional family, the normal family with mean μ and variance σ^2. Here $\theta = (\mu, \sigma^2)$ and $\hat{\theta} = (\overline{y}, s) = (33.31, 7.642)$, where $\overline{y}$ is the average of the 505 ages, and s is their standard deviation. Fourier did not report the data, but he reported the average $\overline{y} = 33.31$ years, and we can calculate the standard deviation $s = 7.642$ years from the error probabilities he reported.

Writing $l(\theta)$ for $\ln(L(\theta))$, we have

$$l(\mu, \sigma^2) = n\left(\ln(s) - \ln(\sigma) - \frac{s^2 + (\overline{y} - \mu)^2}{2\sigma^2} + \frac{1}{2}\right), \quad (3.3)$$

where $n = 505$. The values of (μ, σ^2) for which (3.3) exceeds $\ln(1/2)$ constitute the very good description range for (μ, σ^2), those for which it exceeds $\ln(1/5)$ the good range, those for which it exceeds $\ln(1/15)$ the satisfactory range.

Following Fourier, we are interested only in description ranges for μ. So our question is what values of μ are included

[11] See [9, p. 198]. The annual reports issued by the bureau during Fourier's tenure list no editor on their cover pages. Fourier was no doubt primarily responsible for editing them, and I identify him as the editor in the references [23, 24].

Table 1. Fourier's error probabilities.

probability	1/2	1/20	1/200	1/2000	1/20000
error	±2.7528	±7.9932	±11.4516	±14.2044	±16.5480
level	50%	95%	99.5%	99.95%	99.995%
interval	(33.1, 33.5)	(32.6, 34.0)	(32.4, 34.3)	(32.1, 34.5)	(31.9, 34.7)

The first two lines give the probabilities for errors *in months* that Fourier calculated for his estimate of 33.31 *years*. He gave 1/20, for example, as the probability of the estimate erring by more than 11.4516 months. The second two lines translate this information into confidence intervals in years. The 95% interval, for example, is 32.6 to 34.0 years.

Table 2. Description ranges for the masculine generation in Fourier's study population, calculated from (3.5).

C		description range
2	Very good	$(32.9, 33.7)$
5	Good	$(32.7, 33.9)$
15	Satisfactory	$(32.5, 34.1)$

in these three description ranges. So we maximize (3.3) for each μ, by setting σ^2 equal to $s^2 + (\overline{y} - \mu)^2$. This gives

$$l(\mu, s^2+(\overline{y}-\mu)^2) = n\left(\ln(s) - \frac{1}{2}\ln(s^2 + (\overline{y} - \mu)^2)\right). \quad (3.4)$$

This is greater than $\ln(1/C)$ when μ is in the interval

$$\overline{y} \pm s\sqrt{C^{2/n} - 1}. \quad (3.5)$$

Here we have obtained the log-likelihood (3.4) for the parameter of interest μ by maximizing over the unwanted parameter σ^2. In inferential likelihood theory, the result of such a maximization is sometimes called a *profile likelihood*. The inferential use of profile likelihoods is hard to justify and sometimes misleading [45, p. 159], but their game-theoretic descriptive use is unobjectionable. We want to know how well each value of μ can perform in the betting competition. Its performance depends on the value of σ^2 it is paired with. We pair it with the value of σ^2 that enables it to do its best to see how well it can perform.

Table 2 uses (3.5) to calculate description ranges. Comparing it with Table 1, we see that the width of Fourier's 95% confidence interval is between that of our good description range and our satisfactory description range. The relation as n increases between confidence intervals and the description ranges given by (3.5) has been studied by [10] and [42].

3.2.3 A Fictional Survey of Perceptions

Some organizations in the United States have recently surveyed their employees about perceptions of discrimination. To avoid the complexities involved in real examples, consider the following fictional example.

An organization wants to know whether its employees of different genders and racial identities differ systematically in their perception of discrimination. Most of the employees respond to a survey asking whether they have suffered discrimination because of their gender or race. The employees saying yes are distributed as shown in Table 3.

According to the usual test for the difference between two proportions, the difference between the rows (male vs female) and the difference between the columns (BIPOC vs White) are both statistically significant at the 5% level. But the 20 percentage-point difference between BIPOC males and BIPOC females is not, as its standard error is

$$\sqrt{\frac{1}{3}\frac{2}{3}\left(\frac{1}{20} + \frac{1}{10}\right)} \approx 0.18 = 18 \text{ percentage points.}$$

These simple significance tests seem informative. The differences declared statistically significant seem general enough to be regarded as features of the organization, but we hesitate to say this about the difference declared not statistically significant.

Yet the theory of significance testing does not fit the occasion. Have the individuals in the study (or their responses to the survey) been chosen at random from some larger population? Certainly not. For anyone who has been inside an organization long enough to see its employees come and go, seeing or guessing the reasons, the idea that they constitute a random sample is phantasmagoria. Nor can we agree that their responses are independent with respect to some "data-generating mechanism". Many of them see the same media and talk with each other.

If we took the theory of significance testing seriously for this example, we would also worry about multiple testing. The 5% error rate we claim for our tests is valid under the theory's assumptions only when we make just a single comparison. We have made three comparisons and might make more.

The theory's assumptions are not met, and we have abused the theory. But there is a larger issue. The theory is irrelevant from the outset, because its goals are irrelevant. The organization did not undertake the survey in order to make inferences about a larger or a different population or about some data-generating mechanism. The organization wanted only to know about itself. It wanted to know how its employees' perceptions vary with gender and race. This calls for a descriptive analysis.

For a descriptive analysis, we can use the obvious forecasting family, in which each cell in the 2×2 table has its

Table 3. A fictional study.

	Female	Male	Totals
BIPOC	$\frac{8}{10} = 80\%$	$\frac{12}{20} = 60\%$	$\frac{20}{30} \approx 67\%$
White	$\frac{20}{50} = 40\%$	$\frac{20}{120} \approx 17\%$	$\frac{40}{170} \approx 24\%$
Totals	$\frac{28}{60} \approx 47\%$	$\frac{32}{140} \approx 23\%$	$\frac{60}{200} = 30\%$

Numbers and proportions of positive responses, in a fictional study of the employees of a fictional organization, to the question whether one has experienced discrimination in the organization as the result of one's identity. Here BIPOC means Black, indigenous, and people of color.

own forecast:

$$\theta = (\theta_{\mathrm{bf}}, \theta_{\mathrm{bm}}, \theta_{\mathrm{wf}}, \theta_{\mathrm{wm}}),$$

where θ_{bf} is the forecast that a BIPOC female will say yes to the survey, etc. According to the data in Table 3,

$$\hat{\theta} = \left(\frac{8}{10}, \frac{12}{20}, \frac{20}{50}, \frac{20}{120}\right),$$

and

$$L(\theta) = \left(\frac{\theta_{\mathrm{bf}}}{4/5}\right)^{8} \left(\frac{(1-\theta_{\mathrm{bf}})}{1/5}\right)^{2} \left(\frac{\theta_{\mathrm{bm}}}{3/5}\right)^{12} \left(\frac{(1-\theta_{\mathrm{bm}})}{2/5}\right)^{8} \left(\frac{\theta_{\mathrm{wf}}}{2/5}\right)^{20} \left(\frac{(1-\theta_{\mathrm{wf}})}{3/5}\right)^{30} \left(\frac{\theta_{\mathrm{wm}}}{1/6}\right)^{20} \left(\frac{(1-\theta_{\mathrm{wm}})}{5/6}\right)^{100}.$$

We found earlier that the 20 percentage-point difference between BIPOC males and BIPOC females is not statistically significant. In this descriptive analysis, the question can be reframed this way: what differences between BIPOC males and BIPOC females within the study population are forecast by very good forecasters? We can answer the question by looking at all the $\theta = (\theta_{\mathrm{bf}}, \theta_{\mathrm{bm}}, \theta_{\mathrm{wf}}, \theta_{\mathrm{wm}})$ that rank as very good forecasters by having a value of $L(\theta)$ greater than $1/2$ and finding the range of their values for $\theta_{\mathrm{bf}} - \theta_{\mathrm{bm}}$. The range is from a little more than 0 to about 0.4. We can say that there are very good forecasters who give nearly the same forecasts for the two groups.

When the individuals responding to a yes-no survey are categorized in more than one way, or when other data is collected about them, we may prefer to use a more sophisticated forecasting family, such as logistic regression. The logic will remain the same. For particular interesting values of the forecasting variables, we can calculate the range of forecasts given by good forecasters. We can similarly calculate ranges for odds ratios. The computations involved are not trivial, but software environments adequate to the task would not be need to be more complex for the user than those that now use logistic regression for nominally inferential analyses.

The descriptive approach can be compared with the inferential approach used in 2016 by the University of Michigan's Diversity, Equity & Inclusion Initiative. Michigan sought inferential legitimacy by using random samples. As they explained in their report on the faculty survey [41, p. 6],

> Given the large faculty population at the University of Michigan, this study used a sample survey approach rather than a census of all faculty. A carefully selected sample, with randomization, allows researchers to make scientifically based inferences to the population as a whole.

The second sentence of this quotation raises the question, unanswered by the authors, of how they would have made scientifically based inferences had they performed a whole census. How would they then have decided which differences were meaningful?

In any case, the authors chose 1,500 out of 6,700 faculty members at random to complete the survey. The survey results were then analyzed using logistic regression, and a number of differences were observed to be statistically significant. It was found, for example, that female faculty were 130% more likely to feel discriminated against than male faculty (i.e., the odds ratio for a positive response to the question was equal to 2.3 and significantly different from 1). The results of the survey were clearly meaningful, but the inferential logic is problematic. As David A. Freedman [27] has shown, randomization probabilities do not justify logistic regression. Our descriptive theory is not affected by this problem and is just as applicable to a complete census as to a random or non-random sample.

3.3 Discussion

There is no need to document here the persistence and prevalence over the past two centuries of the use of inferential methods with non-random samples. These abuses have been repeatedly documented and deplored.[12] There have also been numerous efforts to promote descriptive alternatives, but they have gained limited traction in the natural and social sciences. Why can we hope that the proposal to measure relative descriptive success by betting success might fare better? Because, as I have been arguing, our culture already knows something about interpreting betting success.

What are the descriptive alternatives?

3.3.1 Refuse to Calculate Precision

One obvious alternative is to fit models — i.e., calculate descriptive statistics such as means, standard deviations, and regression coefficients — while de-emphasizing the question of their precision and refraining from calculating significance tests or confidence intervals. This has often been the practice in fields, such as geodesy, that have been more influenced by the Gaussian tradition than by the Laplacean tradition.[13] A surveyor's customers, for example, need a boundary line, not a confidence band. Efforts to

[12] See [28, pp. 212–217] for a particularly concise review. See also [1, 6, 9, 26, 39, 40, 48].

[13] Marie-Françoise Jozeau has documented the competition of the two traditions in geodesy [36, 47].

establish the same practice in the social sciences have usually had limited and ephemeral impact. During much of the 20th century, some sociologists avoided using significance testing except for random samples, but this avoidance did not endure. In a study of the use of significance testing in two prominent sociology journals from 1935 to 2000 [38], Erin Leahey found a shift around 1975 among authors who had data on an entire population. Before that date, some of these authors declined to use significance tests, afterwards few did.

In 2004, in his *Regression Analysis: A Constructive Critique* [5], the sociologist Richard A. Berk gave three cheers for a purely descriptive use of multiple regression: estimate the regression coefficients, but do not calculate p-values or confidence intervals. The book has been widely cited, but so far as I know few researchers have followed this advice. What use is a regression coefficient if you have no sense of its precision?

John Tukey's advocacy of data analysis was another effort to shift attention from inference to description [15]. But Tukey's distinction between exploratory data analysis and confirmatory data analysis undermines the effort to promote description to a goal in itself. The terminology suggests that we explore in order to discover what we will then try to confirm.

3.3.2 Reinterpreting Inferential Quantities

In 1995, Richard A. Berk and his colleagues Bruce Western and Robert E. Weiss proposed that the neo-Bayesian philosophy could justify statistical modeling for entire populations [7]. I have not been able to understand their argument.

Another notable effort to make inferential tools descriptive was mounted earlier by [25]. In the simplest cases, their proposal involves permuting residuals and interpreting a p-value as a measure of how unusual the actual data is in the population of alternative data thus generated. Along with many other mathematical statisticians, I was very intrigued by this proposal when it appeared, but I always found its intuition elusive.

The most common justification of using inferential tools with non-random samples is, of course, the argument that assumptions are never exactly satisfied. George Box's slogan, "all models are wrong but some are useful" is evoked to justify calculating p-values and confidence intervals. This leaves unanswered, however, the question of what we are being asked to have confidence in. Any claim that the calculations are mere description is contradicted by the language being used.

3.3.3 Other Justifications for the Likelihood Ratio?

I have proposed scoring relative success by the likelihood ratio $P_1(y)/P_2(y)$. This is hardly a new idea; it goes back at least to Laplace. What this paper adds is a betting interpretation. The ratio is the factor by which algorithm 1 multiplies its capital betting against algorithm 2.

The ratio $P_1(y)/P_2(y)$ can also be understood in terms of information theory. John L. Kelly Jr. entitled his 1956 paper "A new interpretation of information rate"; for him Kelly betting was merely an interesting sidelight to information theory. The connection has been elaborated using the idea of minimum description length [44, 30]. But these ideas too have failed to gain traction in the natural and social sciences, and we cannot expect that they will. As beautiful and simple as they are for mathematicians, the ideas are not part of our wider culture.

3.3.4 Other Loss Functions?

If we set the betting interpretation aside and change the scale by taking logarithms, $P_1(y)/P_2(y)$ becomes $\ln(P_1(y))-\ln(P_2(y))$. As this makes clear, we are comparing the two forecasts using log loss. There are many other loss functions. Why shouldn't we give others equal attention?

The answer, of course, is that we are not setting the betting interpretation aside. A vast amount of work has been done, in decision theory and more recently in machine learning, on evaluating predictions using loss functions. But in most fields of natural and social science, this work has gained little to no traction. I think this is because the scales of measurement have no meaning to the scientists and their larger public. Even though the cutoffs and other conventions we must use will be relatively arbitrary, the idea of measuring relative success in prediction by success in betting and resistance against betting does have meaning for scientists and the publics with whom they want to communicate.

ACKNOWLEDGEMENTS

I have had helpful discussions on these topics with countless colleagues, most recently with Marshall Abrams, John Aldrich, Gert de Cooman, Harry Crane, Nancy DiTomaso, Ruobin Gong, Peter Grünwald, Zev Hirsch, Wouter Koolen, Kitae Kum, Barry Loewer, Aaditya Ramdas, Judith ter Schure, Vladimir Vovk, Ruodu Wang, Sandy Zabell, and Xueyin Zhang. I have also benefited from comments by the editors and several anonymous referees. On the topic of optional continuation, the workshop *Safe, Anytime-Valid Inference (SAVI) and Game-theoretic Statistics* (May 30–June 3, 2022 in Eindhoven, Netherlands) was especially useful.

Accepted 2 December 2023

REFERENCES

[1] Alexander, N. (2015). What's more general than a whole population? *Emerging Themes in Epidemiology* **12**(1) 1–5.

[2] Augustin, T., Coolen, F. P. A., de Cooman, G. and Troffaes, M. C. M., eds. (2014) *Introduction to Imprecise Probabilities*. Wiley. https://doi.org/10.1002/9781118763117. MR3236913

[3] Barnard, G. (1947). Review of *Sequential Analysis* by Abraham Wald. *Journal of the American Statistical Association* **42**(240) 658–665. MR0020764

[4] Barnard, G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society B* **11**(2) 115–149. MR0034983

[5] Berk, R. A. (2004) *Regression Analysis: A Constructive Critique*. SAGE.

[6] Berk, R. A. and Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, 2nd ed. 235–254. Aldine de Gruyter, Berlin.

[7] Berk, R. A., Western, B. and Weiss, R. E. (1995). Statistical inference for apparent populations. *Sociological Methodology* **25** 421–458.

[8] Breiman, L. (1961). Optimal gambling systems for favorable games. In *Fourth Berkeley Symposium on Probability and Mathematical Statistics* (J. Neyman, ed.) **1** 65–78. University of California Press. MR0135630

[9] Bru, B., Bru, M.-F. and Bienaymé, O. (1997). La statistique critiquée par le calcul des probabilités: Deux manuscrits inédits d'Irenée Jules Bienaymé. *Revue d'Histoire des Mathématiques* **3** 137–239. MR1620388

[10] Cahusac, P. (2020) *Evidence-Based Statistics: An Introduction to the Evidential Approach — from Likelihood Principle to Statistical Practice*. Wiley.

[11] Darling, D. A. and Robbins, H. (1967). Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America* **58**(1) 66–68. https://doi.org/10.1073/pnas.58.1.66. MR0215406

[12] Dawid, A. P. (1984). Present position and potential developments: Some personal views. Statistical theory, the prequential approach. *Journal of the Royal Statistical Society: Series A* **147**(2) 278–290. https://doi.org/10.2307/2981683. MR0763811

[13] Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference. *Journal of the Royal Statistical Society: Series B* **53**(1) 79–109. MR1094276

[14] Diaconis, P. and Skyrms, B. (2018) *Ten Great Ideas about Chance*. Princeton. MR3702017

[15] Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics* **26**(4) 745–766. https://doi.org/10.1080/10618600.2017.1384734. MR3765335

[16] Doob, J. L. (1953) *Stochastic Processes*. Wiley, New York. MR0058896

[17] Edwards, A. W. F. (1972) *Likelihood: An Account of the Statistical Concept of* Likelihood *and its Application to Scientific Inference*. Cambridge. MR0348869

[18] Ethier, S. (2010) *The Doctrine of Chances: Probabilistic Aspects of Gambling*. Springer, Berlin. https://doi.org/10.1007/978-3-540-78783-9. MR2656351

[19] Feller, W. (1950) *An Introduction to Probability Theory and Its Applications*, 1st ed. Wiley, New York. MR0038583

[20] Feller, W. K. (1940). Statistical aspects of ESP. *The Journal of Parapsychology* **4**(2) 271–298. MR0004461

[21] Fisher, R. A. (1952). Sequential experimentation. *Biometrics* **8** 183–187.

[22] Fisher, R. A. (1956) *Statistical Methods and Scientific Inference*. Hafner. Later editions in 1959 and 1973. MR0131909

[23] Fourier, J. (1826). Mémoire sur les résultats moyens déduits d'un grand nombre d'observations. In *Recherches Statistiques sur la Ville de Paris et le dÉpartement de la Seine* (J. Fourier, ed.). Imprimerie Royale, Paris.

[24] Fourier, J. (1829). Second mémoire sur les résultats moyens et sur les erreurs des mesures. In *Recherches statistiques sur la ville de Paris et le département de la Seine* (J. Fourier, ed.) Imprimerie royale, Paris.

[25] Freedman, D. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business and Economic Statistics* **1** 292–298.

[26] Freedman, D. A. (1991). Statistical models and shoe leather. *Sociological Methodology (with discussion)* **21** 291–358.

[27] Freedman, D. A. (2008). Randomization does not justify logistic regression. *Statistical Science* **23**(2) 237–249. https://doi.org/10.1214/08-STS262. MR2516822

[28] Freedman, D. A. (2009) *Statistical Models: Theory and Practice, Revised Edition*. Cambridge. https://doi.org/10.1017/CBO9780511815867. MR2489600

[29] Greenwood, J. A. (1938). An empirical investigation of some sampling problems. *The Journal of Parapsychology* **3**(2) 222–230.

[30] Grünwald, P. (2007) *The Minimum Description Length Principle*. MIT.

[31] Grünwald, P., de Heide, R. and Koolen, W. M. (2020). Safe testing. In *2020 Information Theory and Applications Workshop (ITA)* 1–54. IEEE.

[32] Grünwald, P., Ly, A., Pérez-Ortiz, M. and ter Schure, J. (2021). The safe logrank test: Error control under optional stopping, continuation and prior misspecification. *Proceedings of Machine Learning Research* **146** 107–117.

[33] Hendriksen, A. A. (2017). *Betting as an alternative to p-values*. Master's thesis, University of Leiden, under the direction of Peter Grünwald.

[34] Hexter, R., Pfuntner, L. and Haynes, J., eds. (2020) *Appendix Ovidiana: Latin Poems Ascribed to Ovid in the Middle Ages*. Harvard.

[35] Howson, C. and Urbach, P. (2006) *Scientific Reasoning: The Bayesian Approach*, Third ed. Open Court.

[36] Jozeau, M.-F. Géodésie au XIXème Siècle: De l'hégémonie française à l'hégémonie allemande. Regards belges (1997). PhD thesis, Université Denis Diderot Paris VII, Paris.

[37] Kelly Jr., J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal* **35**(4) 917–926. https://doi.org/10.1002/j.1538-7305.1956.tb03809.x. MR0090494

[38] Leahey, E. (2005). Alphas and asterisks: The development of statistical significance testing standards in sociology. *Social Forces* **84**(1) 1–24.

[39] Mason, W. M. (1991). Freedman is right as far as he goes, but there is more, and it's worse. Statisticians could help. *Sociological Methodology* **21** 337–351.

[40] Matthews, J. R. (1995) *Quantification and the Quest for Medical Certainty*. Princeton.

[41] Office of Diversity, U. o. M. Equity & Inclusion (2017). *Results of the 2016 University of Michigan Faculty Campus Climate Survey on Diversity, Equity & Inclusion*. https://diversity.umich.edu/wp-content/uploads/2017/11/DEI-FACULTY-REPORT-FINAL.pdf

[42] Pawel, S., Ly, A. and Wagenmakers, E.-J. (2023). Evidential calibration of confidence intervals. *American Statistician*.

[43] Ramdas, A., Grünwald, P., Vovk, V. and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science* **38**(4) 576–601. https://doi.org/10.1214/23-sts894. MR4665027

[44] Rissanen, J. (1989) *Stochastic Complexity in Statistical Inquiry*. World Scientific. MR1082556

[45] Royall, R. M. (1997) *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall. MR1629481

[46] Shafer, G. (2007). From Cournot's principle to market efficiency. In *Augustin Cournot: Modelling Economics* (J.-P. Touffut, ed.) 55–95. Edward Elgar.

[47] Shafer, G. (2019). *On the nineteenth-century origins of significance testing and p-hacking*. Working paper 55, Game-Theoretic Probability Project.

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

[49] Shafer, G. (2020). *Game-theoretic foundations for statistical testing and imprecise probabilities*. SIPTA School 2020/2021, Slides and videos.

[50] Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication (with discussion). *Journal of the Royal Statistical Society: Series A* **184**(2) 407–478. https://doi.org/10.1111/rssa.12647. MR4255905

[51] Shafer, G. (2022). *That's what all the old guys said: The many faces of Cournot's principle*. Working Paper 60,

Game-Theoretic Probability Project.
[52] Shafer, G. and Vovk, V. (2001) *Probability and Finance: It's Only a Game*. Wiley, New York. https://doi.org/10.1002/0471249696. MR1852450
[53] Shafer, G. and Vovk, V. (2006). The sources of Kolmogorov's *Grundbegriffe*. *Statistical Science* **21**(1) 70–98. See also Working Paper 4, Game-Theoretic Probability Project.
[54] Shafer, G. and Vovk, V. (2019) *Game-Theoretic Foundations for Probability and Finance*. Wiley, New York. https://doi.org/10.1002/0471249696. MR1852450
[55] Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011). Test martingales, Bayes factors and p-values. *Statistical Science* **26**(1) 84–101. https://doi.org/10.1214/10-STS347. MR2849911
[56] Troffaes, M. C. M. and de Cooman, G. (2014) *Lower Previsions*. Wiley. https://doi.org/10.1002/9781118762622. MR3222242
[57] Vovk, V. (2023). *The diachronic Baysian*. Working paper 64, The Game-Theoretic Probability Project, http://probabilityandfinance.com.
[58] Vovk, V. and Wang, R. (2021). E-values: Calibration, combination, and applications. *Annals of Statistics* **49**(3) 1736–1754. https://doi.org/10.1214/20-aos2020. MR4298879
[59] Wald, A. (1947) *Sequential Analysis*. Wiley, New York. MR0020764
[60] Ziemba, W. T. (2015). A response to Professor Paul A. Samuelson's objections to Kelly capital growth investing. *The Journal of Portfolio Management* **42**(1) 153–167.

Glenn Shafer. Rutgers Business School and Department of Statistics, 1 Washington Park, Newark, New Jersey 07102, USA. E-mail address: gshafer@business.rutgers.edu