What is Game-Theoretic Statistics?

Glenn Shafer

March 10, 2025

Abstract

Statistical modelling is the art of connecting abstract models with scientific or practical questions, so that the models' probabilities can throw light on these questions. Standard statistical theory summarizes the results with p-values. Unfortunately, p-values are widely misunderstood. Game-theoretic statistics replaces them with e-values, which can be interpreted in terms of a game. The model makes a forecast, the statistician bets against it, and the e-value is the factor by which she multiplies her stake.

1 Testing by betting

If you put one euro on the table and use it to make a fair bet, what do you expect to get back? If the bet is all-or-nothing, as when the bet is settled by tossing a coin, you get back 2 or 0. Other bets may have more than two possible payoffs. But the concept of *expected value* provides one way of summarizing your expectation. The expected value is found by multiplying each amount you might get back by its probability and adding. Mathematicians say that a bet is *fair* if the expected value is equal to the amount you pay — i.e., the amount you put on the table. So the all-or-nothing bet is fair if the probabilities of 2 and 0 are both 1/2, so that

$$\frac{1}{2}2 + \frac{1}{2}0 = 1.$$

In this case we often say that the coin itself is fair.

Here is a more complicated example. Suppose you bet on how many tosses of a fair coin will be needed to get the first head. Probability theory tells us that the probability that the first head will appear on the nth toss is $1/2^n$. Suppose your bet requires you to pay one euro and pays back

$$\frac{9^{n-1}}{5^n} \tag{1}$$

euros if the first head appears on the nth toss. This bet is fair because

$$\sum_{n=1}^{\infty} \frac{1}{2^n} \frac{9^{n-1}}{5^n} = \sum_{n=1}^{\infty} \frac{9^{n-1}}{10^n} = \frac{1}{10} \sum_{n=1}^{\infty} \left(\frac{9}{10}\right)^{n-1} = 1.$$

It might also have a very surprising outcome. Suppose the first head appears on the 100th toss. Then (1) tell us that the payback from your investment of one euro is more, in euros, than 3 followed by 24 zeros. This shows that the bet is entirely imaginary, because the wealth of the entire world is very, very tiny in comparison. But not seeing heads on the first 99 tosses gives us strong reason to think that the probability distribution that assigns probability $1/2^n$ to the first head appearing on the nth toss is not a good description of the world. *It is not a good forecast of what will happen.*

Why would a statistician ever think about the imaginary bet (1)? She would have good reason to think about it if she thought that the probability of heads on each trial is 1/10 instead of 1/2. In this case, her probability for seeing $n-1$ tails followed by a head is

$$\left(\frac{9}{10}\right)^{n-1}\frac{1}{10}, \tag{2}$$

and (1) is the ratio of (2) to $1/2^n$. Statisticians generally agree that such a ratio, called the *likelihood ratio*, is a good measure of the forecasting success of one probability distribution relative to another.

Waiting so long for the first head would also make our statistician (let's call her *Statistician* with a capital S) question her own probabilities, because they assign a probability less than 0.000033 to the first 99 tosses being tails. But a less extreme outcome might confirm Statistician's opinion. For example, if the first head appears on the 10th toss, then (1) is approximately 40 and exactly equal, on a logarithmic scale, to the expected value attributed to (1) by Statistician's probability distribution.[1] Multiplying your capital by 40 is not like multiplying it by 3 followed by a hundred thousand zeros. But it is already astonishingly strong as evidence against the probabilities $1/2^n$.

2 Test martingales

Instead of imagining a single bet at the outset, Statistician might imagine a sequence of bets, one on each toss.

1. First she pays 1, getting back 9/5 if the first toss comes out tails and 1/5 if it comes out heads.

2. She stops betting if the toss came out heads. If it came out tails, she pays the 9/5 she now has to get $(9/5)^2$ if the first toss comes out tails and $(9/5)(1/5)$ if it comes our heads.

3. And so on. If she has all tails after k tosses, she bets her capital at that point, $(9/5)^k$, to get back $(9/5)^{k+1}$ if the $(k+1)$st toss comes out tails and $(9/5)^k(1/5)$ if it comes out heads.

[1] If you have studied probability theory, then you know that Statistician's probability distribution is a geometric distribution with expected value 10, and you can verify that the expected value of the logarithm of (1) under this distribution is equal to the logarithm of its value when $n = 10$.

The net result will be the same as the first bet we studied: Statistician begins with 1 and ends up with $(9/5)^{n-1}(1/5)$, where n is the number of tosses to get a head.

This type of nonnegative process, where the results of successive bets are compounded, is called a *test martingale*.[2] It turns out that any bet on a sequence of outcomes that begins with unit capital and has a nonnegative payoff can be obtained by a test martingale.

3 Two ways of testing: e-values and p-values

The result of a bet that has a nonnegative payoff with expected value or less 1 under the hypothesis or hypotheses being tested is called an *e-value*. As we have just seen, a test martingale is the result of multiplying successive e-values.[3]

In order to use e-values (and test martingales) for statistical inference, Statistician must adopt the *principle* that a large e-value, for a bet chosen in advance, discredits the hypothesis. It should be kept in mind that this is a principle, not a theorem. How large is large? This again is a matter of judgement. Some statisticians have established conventions about how to describe the degree of discredit associated with different e-values. One convention is that 4 is serious evidence, 10 is strong evidence, and multiples of 10 are very convincing. The value of such a convention depends on the context, however. A large e-value can be taken seriously only if it is based on plausible alternatives that Statistician has good reason to consider. Otherwise, it can easily be dismissed as a lucky stab in the dark or even as a likely mistake.[4]

The evidentiary role of an e-value is analogous to that of a *p-value*. We used a p-value in §1, when we observed that Statistician's own hypothesis assigns a probability of only about 3 in a hundred thousand to the first head coming only on the 100th toss or later. Statisticians usually define a p-value by first choosing a test statistic, or function of the data; a hypothesis's p-value is its probability for the test statistic being as large or larger than it actually turned out to be.

Statisticians have used p-values for two hundred years, and they remain more widely used than e-values by a huge margin. But e-values have several advantages. As we have seen, we can combine evidence from successive observations by multiplying e-values. The average of two or more e-values, possibly weighted, is also an e-value, and this provides a way of combining evidence based on different alternatives to the hypothesis being tested. If one of the bets produces

[2]For centuries, betting strategies in casinos were called martingales. Now mathematicians use the name for capital processes that result from betting strategies [6].

[3]Jean Ville studied nonnegative martingales starting with unit capital in the 1930s [10]. Herbert Robbins and his collaborators used composite nonnegative supermartingales in statistics beginning in the 1960s [1]. Vladimir Vovk emphasized the direct interpretation of high values of a test martingale in the early 1990s [11]. The simpler concept of an e-value emerged in the late 2010s [2, 3, 12, 14]. The name *e-value* is due to Vovk and Ruodu Wang [12]. Unfortunately, there are several other established uses for the name in various specialized fields of statistics. An alternative name for our concept of an e-value is *betting score* [8].

[4]In addition to testing, there are other applications of e-values. In some applications, modest e-values are useful [4]. In some others, only very large e-values are considered [13].

a sufficiently large e-value, its average will also be large. Combining successive observations or distinct tests is possible for p-values but more awkward. One way to combine p-values is to convert them into e-values. There are many functions that do this; the simplest may be $1/\sqrt{p} - 1$. It converts the p-value 0.05, conventionally considered serious evidence, to the e-value 3.47.

When we simultaneously test more than one hypothesis about the same outcomes, we say that we are testing a *composite hypothesis*. An e-value for a composite hypothesis is a payoff that has expected value 1 or less under each of the hypotheses. A p-value is the maximum or supremum of the different hypotheses' p-values using the same test statistics.

4 Free play and a fundamental principle

If Statistician was testing the fairness of her coin using the martingale we described in §2, then she saw this sequence of cumulative e-values:

$$\frac{9}{5},\ \left(\frac{9}{5}\right)^2,\ \left(\frac{9}{5}\right)^3,\ \left(\frac{9}{5}\right)^4,\ \left(\frac{9}{5}\right)^5,\ \left(\frac{9}{5}\right)^6,\ \left(\frac{9}{5}\right)^7,\ \left(\frac{9}{5}\right)^8,\ \left(\frac{9}{5}\right)^9,\ \left(\frac{9}{5}\right)^9,\ \left(\frac{1}{5}\right)$$

But after the fourth toss, she already had an e-value of $(\frac{9}{5})^4$, which is greater than 10. At that point, she might have felt that this was all the evidence against the coin's fairness that she needed. So she might have stopped. Is this legitimate? Statisticians agree that it is, but her right to stop and count $(\frac{9}{5})^4$ as evidence is again a *principle*, not a theorem. It has been called the *principle of optional stopping*.

Here is a more general principle that is equally reasonable.

> **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.

This principle not only allows Statistician to stop when she wants, it also allows her to begin testing with no plan for when to stop or how to bet on later rounds if she does not stop. It even allows her to change the experiment that produces the outcomes she is betting on. In short, it allows *free continuation*.

5 Statistician's logarithmic utility

Human reactions to many stimuli are logarithmic. This also seems true of e-values. The difference between the two e-values 5 and 25 seems important, while the difference between 105 and 125 seems negligible. If Statistician has her own probabilities, we can formalize this intuition by assuming that she also has a logarithmic utility function, in the sense that she chooses how to bet so as to maximize her expected value of the logarithm of her e-value.[5]

[5]The assumption that a decision maker maximizes subjective expected utility is widely made in decision theory and economics.

A bit of notation may help here. Statistician is betting against a probability distribution P for some outcome Y. She has her own probability distribution Q for Y. Her bet will be a nonnegative function S of Y that is assigned expected value 1 or less by P; we can write this condition as $\mathbf{E}_P(S(Y)) \leq 1$. Maximizing her expected utility means choosing S to maximize $\mathbf{E}_Q(\ln(S(Y)))$. This was advocated by John L. Kelly, Jr., in 1956 [5], and we call it *Kelly betting*.

This maximization problem is not mathematically difficult; the maximum is achieved when S is the likelihood ratio: $S = Q(Y)/P(Y)$.[6] So maximizing expected logarithmic payoff is consistent with the e-value (1) in the problem where a statistician has the probabilities (2) for when the first head will appear.

An important advantage of maximizing the expected value of the logarithm of the payoff follows directly from the fact that Statistician is multiplying e-values. This means adding their logarithms. By maximizing the logarithm on each round, she does all she can on that round towards maximizing the sum of the logarithms. If she chose her bet on a particular round to maximize the expected value of the payoff rather than its logarithm, she might lose all her capital. And then she would not be able to bet on the next round.

6 Statistical modelling

E-values can be used to test forecasters who give probabilities for outcomes that we will observe. They can also be used to make inferences about facts that we will never observe and questions that we do not expect to settle with certainty and precision. What is the mass of Jupiter? Does a particular medical treatment do any good? To address such a question with e-values, Statistician must make an argument for connecting a probability distribution with the question. Statistical modelling is the art of making such arguments.

One classical example of statistical modelling is the theory of errors developed by Laplace and Gauss. Suppose Statistician thinks that a normal probability distribution with mean zero and variance one (the familiar "bell-shaped curve" of statistics) has described past errors of her measuring instrument reasonably well. She argues that this is reason enough to adopt it as her forecast of the error it will make measuring a quantity μ. This is not really a forecast, as she will not observe this error. But she can imagine someone (let's call him Forecaster) who knows μ and therefore can make the forecast and observe the error, and someone else (let's call him Skeptic) who also knows μ and bets against the forecast. Although Statistician does not see this imaginary betting, she sees the measurement and so can calculate how well Skeptic did as a function of μ. As she thinks the normal distribution is a good forecast, she expects that Skeptic will not multiply his capital by a lot, and this will give her guidance about the value of μ.

The division of labor between Forecaster and Skeptic allows us to understand more clearly some of the more complicated arguments statisticians make. In the

[6] For a simple proof, see [8]. For extensions to composite hypotheses, where the maximization can be difficult, see [7].

case of randomized medical experiments, for example, Forecaster can rely on the probabilities involved in the random assignment of treatments, while Skeptic can use Statistician's more subjective beliefs about how different treatments might work with particular patients.

7 Game-theoretic probability

As a branch of mathematics, game theory studies strategies in fully defined games — games for which the mathematician specifies the players' goals and the extent to which they see other players' moves. In real games, players might also have private information and might acquire more private information as play proceeds, but in the mathematical theory a player's strategy can use only the previous moves in the game that the player sees.

We leave the realm of pure mathematics when we allow Forecaster and Skeptic (or Statistician, who is pulling their strings) to play freely instead of following a strategy specified at the outset of play. We also sometimes find it useful to introduce other players who can play freely, including Nature, who announces things unknown to Statistician, like μ in our example, and Experimenter, who may change the experiment for the next round of play or announce auxiliary information, as in regression or classification problems.

It is also interesting and useful, however, to re-enter the realm of mathematics by studying strategies for Forecaster and Skeptic. This is *game-theoretic probability.*[7] To fully define the game, we need a player who announces outcomes; we call this player Reality. We can then ask whether Skeptic has a winning strategy for a particular goal.

Consider, for example, this game between Forecaster, Skeptic, and Reality, which continues for 1,000 rounds. Skeptic begins with 1 euro. Forecaster begins each round by announcing a probability for "heads", then Skeptic bets against the forecast, and then Reality says "heads" or "tails". Skeptic must bet in such a way that his current capital (his initial unit capital plus any winnings minus any losses) never becomes negative, no matter what Reality does. Write $\overline{p}$ for the average of Forecaster's 1,000 probabilities, $\mathcal{K}$ for Skeptic's capital at the end of play, and #heads for number of times Reality says "heads". And set Skeptic this goal:

$$\text{Either } \left| \frac{\#\text{heads}}{1{,}000} - \overline{p} \right| < 0.1 \text{ or else } \mathcal{K} \geq 10. \tag{3}$$

Skeptic has a winning strategy in this game. In other words, Skeptic can guarantee (3) no matter what Forecaster and Reality do.

This is one game-theoretic instance of the many-faceted *law of large numbers.* It illustrates how discrete-time probability theory can be translated into game theory and then generalized. It generalizes Chebyshev's version of the law of large numbers to the case where Forecaster is a free player.

[7] The most thorough study of game-theoretic probability is [9]. Some examples of game-theoretic statistical modelling are provided in Chapter 10.

The game-theoretic law of large numbers supports an argument made by Kelly in 1956 for Skeptic using logarithmic utility. Laws of large numbers, game-theoretic or not, are available only when we are adding. Because the logarithms and their expected values add, reasonable conditions on Skeptic's probabilities will allow him to use a game-theoretic law of large numbers to attribute a high probability to his capital growing, even though Forecaster is a free player. This does not work if he tries to maximize his expected value for his final capital or other functions of it.

8 How to discourage real gambling

Probability theory began as a theory of gambling in games of pure chance. When Jacob Bernoulli, who proved the first law of large numbers, undertook to apply the theory to civil, criminal, and business affairs, he minimized the connection with gambling. Most statisticians have followed his example. We want statistical inference to live in the realm of reason, not in the casino.

None of the mathematical statisticians who have been developing game-theoretic statistics seek to promote gambling. Some even advocate its legal prohibition. But the value of game-theoretic statistics lies in the insights that can be gained by making the betting game between Forecaster and Skeptic explicit, and this makes hiding the connection with gambling impossible. Perhaps we can instead use what we can learn from studying martingales to educate the public about the perils of gambling.

No one has ever bankrupted their family and destroyed their lives by trying to multiply an investment of one euro without risking more. Yet our theory shows that this is the legitimate way to show that you know better than the posted odds, in a casino, at a horse race, or in a financial market. People destroy their lives when, consciously or unconsciously, they play martingales that can become negative, and deeply negative. It has become easier and easier to play such martingales, whether by impulsively making another online sports bet or by speculating in financial securities on margin. It should be the duty of teachers of probability and statistics to teach not only how the law of large numbers works but also how slowly it works, and how easily gamblers and day traders can delude themselves.

Acknowledgments

Game-theoretic statistics is a very new field. Its excitement and promise is matched by the diversity of viewpoints within the community of researchers developing it. This introduction, which emphasizes betting rather than the mathematics of e-values and related concepts, owes the most to my long-time collaborator Vladimir Vovk. I am also grateful for insights provided by Oberwolfach's Workshop 2419b, held in May 2024, and especially by its organizers, Peter Grünwald, Aaditya Ramdas, Ruodu Wang, and Johanna Ziegel. A version

of this note may later appear as an Oberwolfach Snapshot.

References

[1] Donald A. Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America*, 58(1):66–68, 1967.

[2] Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing (with discussion). *Journal of the Royal Statistical Society, Series B*, page to appear, 2024.

[3] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability surveys*, 17:257–317, 2020.

[4] Nikolaos Ignatiadis, Ruodu Wang, and Aaditya Ramdas. E-values as unnormalized weights in multiple testing. *Biometrika*, 111(2):417–439, 2023.

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

[6] Laurent Mazliak and Glenn Shafer, editors. *The Splendors and Miseries of Martingales: Their History from the Casino to Mathematics*. Birkhäuser, 2022.

[7] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

[8] Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication (with discussion). *Journal of the Royal Statistical Society, Series A*, 184(2):407–478, 2021.

[9] Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*. Wiley, New York, 2019.

[10] Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939. (includes Ville's thesis, which is online at Numdam).

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

[12] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

[13] Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society, Series B*, 84(3):822–852, 2022.

[14] Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.