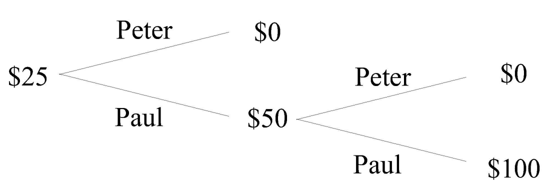
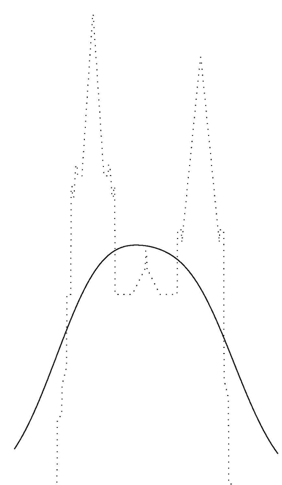
Teaching testing by betting: Towards a syllabus for game-theoretic statistics

Glenn Shafer
Rutgers University
gshafer@business.rutgers.edu
www.glennshafer.com





The Game-Theoretic Probability and Finance Project

Working Paper #62

First posted July 3, 2022. Last revised July 6, 2022.

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

Abstract

This note sketches a syllabus for an introductory course in statistics built around the notion of testing by betting. This sketch is prefaced by an explanation of the author's personal attitudes towards betting and his personal opinions about betting's role in our culture.

The language of betting provides a good way of communicating the results of statistical tests. So I claimed in [19]. Some of the paper's discussants questioned the claim. Can we really teach testing by betting? Is it easier to teach than p-values?

As Judith ter Schure has emphasized, these questions can only be answered experimentally [22]. To see whether we can teach testing by betting, we need to give it a try. In this note, I try to think through how to teach testing by betting to students who may not have already studied p-values or concepts in mathematical statistics.

For clarity, I begin by explaining my personal attitudes towards betting and my personal opinions about its role in our culture. Then I sketch a possible syllabus for a semester-long introductory course in statistics built around testing by betting. I hope to teach such a course at Rutgers, and I welcome advice on how to do so.

1 Cards on the table

I have been writing and talking about betting and probability for forty years, discussing mathematical, philosophical, and historical aspects of the relation between the two. Often, however, my listeners brush the broad picture aside and raise more personal points: their feelings about the moral and cultural aspects of betting and their intuitions about probability.

This experience has convinced me that it is futile to ignore these moral and cultural issues. So I begin by putting my own cards on the table, explaining my own feelings and intuitions and how they have informed my work.

1.1 Opposition to gambling

For whatever reason, I have never been attracted to gambling. But I know that it can become dangerously attractive to many people, and I have always opposed its expansion. In 1974, I voted against legalizing casino gambling in New Jersey. (This proposal failed.) In 1986, I voted against the establishment of a state lottery in Kansas. (This proposal passed.) Most recently, in 2021, I voted against allowing betting on college sports in New Jersey. (This proposal failed.)

I have also never personally ventured into the stock market. As an academic in the United States, I have been fortunate enough to have a well managed pension fund (TIAA-CREF). Other savings have gone into savings accounts, a passive mutual fund (Vanguard), and real estate where I have lived.

1.2 The metaphor of betting in our culture

Growing up, I did not have any experience with gambling. But expressions such as "I bet that ...", "I will bet you a hundred to one", and "Put your money where your mouth is" were a common and even prominent part of the language I learned. I suspect that my experience is not unusual. I suspect that talk about betting has long been one way of expressing belief and confidence for most people in the United States and most other countries.

It is implicit in this sort of talk that success in betting is evidence that the winner is more knowledgeable about the topic of the bets or more clever in using what they know. They are the better forecaster. I want to be careful and tentative here, because systematic observation of actual bets and of the success and failure of particular bettors is surely less widespread in our culture than bold talk about betting. I do not contend that most people have extensive intuitions about the meaning and pitfalls of betting success. I do contend they can develop such intuitions more easily than they can develop intuitions about the meaning and pitfalls of p-values.[1]

1.3 The fallacy of untutored intuition

In his most famous book, *The General Theory of Employment, Interest and Money*, John Maynard Keynes wrote,

> Practical men, who believe themselves to be quite exempt from any intellectual influence, are usually the slaves of some defunct economist. [13, p. 383]

I think I once read some similarly scathing comment by some similarly reknowned author concerning intuition. I cannot remember who it was, so I must speak for myself. I feel that when philosophers (including philosophical statisticians) talk about their intuition, they are talking about some theory they have learned or absorbed, from family, friends, teachers, or books.

This feeling weighs on me whenever someone speaks of intuition about numerical probability as if it were something untutored. Numerical probability is a theoretical invention. Long before the time of Blaise Pascal and Jacob Bernoulli, European intellectuals talked about the probabilities for and against something, but these probabilities were not numbers, and they usually had nothing to do with betting or Pascal's logic of betting.

1.4 Opposition to always assessing evidence in terms of betting odds

My first scholarly publications developed a calculus for combining evidence that used numerical "degrees of belief" (or "evidential support") that did not obey the standard rules of mathematical probability [15]. This calculus came to be called the Dempster-Shafer theory. Many Bayesian critics called it incoherent or irrational, arguing that its degrees of belief would lose money when used as betting rates.

These Bayesian objections partly motivated my mathematical and historical study of the relation between betting and probability. This study has convinced

[1] This point has been argued by Judith ter Schure in her doctoral thesis [22, pp. 13–14].

me that betting games are indeed the historical origin and the only complete justification for the definitions and axioms of standard mathematical probability. But I contend that both Bayesian and non-Bayesian claims for the generality of the standard calculus are flawed for precisely this reason. Not all evidence can be well described or summarized with betting ideas. Sometimes we should not even try to evaluate the strength of given evidence numerically, and sometimes it is useful to make numerical evaluations that do not follow the logic of betting.

1.5 Concerns about morality and frivolity

These are genuine concerns. To address them, we can make these points:

1. Standard mathematical probability was invented as a theory of betting.

2. The role of betting is a secret to no one. Speakers of English understand "odds two to one that ..." as a reference to betting and a statement of numerical probability. Especially today, when gambling is so available online, we need to teach about its dangers rather than averting our gaze.

3. Attempts to give the standard calculus's numerical probabilities a meaning other than betting odds have never been fully satisfactory. We can see this already in Jacob Bernoulli's *Ars conjectandi* (1713), where he generalizes Huygens's theory of chance in games by discerning "equal possibility" in ways other events in life can happen [3, pp. 326–327]. Since the advent of civilization, no doubt, gamblers have thought to generalize their prowess with dice to life's other business, but Bernoulli was the first to attach the word "probability" (*probabilitas* in Latin, *probabilité* in French) to betting mathematics.

4. The interpretation of probabilities as frequencies is misleading, because it makes the independent identically distributed case seem fundamental even though, as Laplace said, it never happens in nature.

5. The frequency interpretation is also mathematically inadequate. As George Barnard asserted [1, p. 263], "Ville showed that any attempt to define randomness along the lines of von Mises-Wald-Church would run against the difficulty, for example, that it was not excluded that the limiting frequency should be approached from above only."

6. To test probabilities, we can bet with pennies or tiny fractions of a penny or, equivalently, just pretend to bet. The size of the bet is irrelevant to the evidential meaning of its success, which depends on how much you multiply the money you risk.

A full understanding of the dangers of gambling will include an understanding of the danger of martingaling [7, 8, 20]. Teaching about martingaling is also an important step in teaching game-theoretic testing, because it clarifies what is meant by "the money you risk".

1.6 What our culture knows about betting success

When two people have different opinions, it is natural to make an even-money bet. This does not mean that anyone thinks the probability is 50–50. Yet it is a fair bet in an obvious sense.

Our culture also has the concept of offering odds to express one's strong opinion. No one will argue, I think, with these ideas:

1. Consistently winning at odds Jane offers is evidence of greater knowledge or insight than Jane.
2. If someone with more knowledge and insight than Jane consistently fails to win betting against Jane's betting offers, then her probabilities are pretty good, and the opponent's greater knowledge and insight are not really relevant to the prediction problem.

Beyond this, however, our culture does not offer a fully developed and widely understood concept of testing by betting that fits the standard probability calculus. We need to develop this concept, and then we need to teach it.

1.7 Teaching Bernoulli and Bayes

While teaching testing by betting, how much attention should we give to established Bernoullian and Bayesian ideas: p-values, likelihood-ratio tests, power, the likelihood principle, priors and posteriors, etc.?

A fair test of testing by betting must, I think, leave competing foundations for statistics to fend for themselves. Twenty-five years ago, Steffen Lauritzen taught me that even when different ways of organizing a mathematical topic are simple on their own terms, relating them to each other is likely to be very complicated.

The topic of p-values is already a muddle, and relating p-values to betting scores only muddies the muddle further. In [19], I argued that the notion of implied target is more useful and coherent than the notion of power, but this argument should not take center stage in a first course in statistics built around testing by betting. In my experience, most people who have already studied and even used statistics do not remember the word "power" and cannot reproduce the definition of "Type II error". Teaching or reteaching these topics would take us too far afield.

Likelihood ratios, confidence intervals, and the averaging of probability distributions are native concepts in game-theoretic statistics. When we encounter them, we must mention that they also have roles in Bernoullian and Bayesian statistics. But detailing those roles will not be our task.

A course that does not teach established concepts and techniques will not prepare students to use those concepts and techniques. The experimental course I am asking us to think about addresses a different purpose and perhaps a different audience. Perhaps it should be an honors course or elective for students more interested in learning something new than in learning methods already widely used.

2 Syllabus for an introduction to game-theoretic statistics

This is a very preliminary syllabus. I seek advice on how to improve it and on resources for teaching each of its lessons.

If I teach a course along these lines at Rutgers. I expect that my greatest difficulty will be in providing computational resources.

2.1 Multiplying the capital you risk

Begin with betting against a single probability (perhaps a probability for rain tomorrow). Then betting against a point prediction (perhaps a prediction of tomorrow's earnings announcement or of the point spread in tomorrow's basketball game). Then betting against a probability distribution by selecting a variable and buying it at its expected value with respect the distribution. Distinguish between *bet* and *betting score* (how much you multiply your money). The bet is a function of the outcome to be observed; the betting score is its realized value.

Multiplying the capital you risk is evidence against the forecaster who advanced the probability, point prediction, or probability distribution. But this evidence needs to be weighed, usually informally, with other evidence, positive or negative. Everyone knows that you might just have been lucky; on this point our culture is already sophisticated enough.

Explain that the factor by which you multiply your capital is invariant to the amount bet. So the amount bet does not matter; it can be so small that the outcome of the bet does not matter to anyone monetarily. Or the betting can just be imagined. But the bet must still be declared in advance, just like a real bet.

You will seldom be able to multiply the capital you risk a lot with a single bet. But you might use the total capital resulting from your bet to make another bet against the same forecaster. Repeating this many times, you may obtain a large betting score.

Give examples with data.

2.2 Kelly and Neyman-Pearson betting

Explain these two ways of selecting a bet if you have your own probabilities for the outcome being predicted. Discuss half-Kelley and and half-Neyman-Pearson as well. Kelley can be used when the forecaster does not offer an entire probability distribution. Neyman-Pearson is less flexible.

If the forecaster offers an entire probability distribution, then your choice of a bet defines an alternative for which the bet is the Kelly bet. The betting score is a likelihood ratio.

How might you choose a bet if you do not have an entire probability distribution for the outcome yourself?

1. You might have some beliefs of your own that fall short of a probability distribution but still give you some hints.

2. You might want to test some feature of the forecast without necessarily having different opinions of your own. You might question, for example, the level of confidence of a forecaster who gives a probability close to zero or one. More generally, you might question the precision of a probability distribution for a numerical outcome. When the forecast is a joint probability distribution for several numerical outcomes, there are many aspects that you might question, such as a high probability for the outcomes to be close or far apart. Etc. If you want to bet against a particular feature of a probability distribution P, say its precision, then you might make a Kelly bet based on an alternative Q that resembles P but is more diffuse. In this case, Q does not represent probabilities of your own.

3. You may simply choose a bet wildly, blindly. In this case, success in multiplying your capital may be less convincing as evidence against the forecaster, because your success is so clearly a matter of luck.

The interpretation of success in testing by betting depends on what the person doing the testing is trying to accomplish. Is she championing probabilities of her own? Is she checking some particular possible defect of the forecaster's probability or probabilities — that the probability is too far from one-half or that the probabilities are too spread out? Or is she just looking randomly for weak spots?

This is an inescapable feature of statistical testing. Without full and honest disclosure by the person doing the testing, others cannot reliably interpret their results.

2.3 Testing probabilities of rain for successive days

Each time you use the capital from the previous bet, never risking more than you have. More precisely, you begin by putting on the table the most your first bet risks. After each bet, you add your gain from that bet, positive or negative. Otherwise, you never put more money on the table, and you never make a bet that risks more than what you have on the table.

The fundamental principle at work here is that you can discredit a forecaster and his probabilities by multiplying your capital by a large factor betting against them. This principle requires that you make the bet against each forecast as soon as the forecaster announces it. But it permits much other flexibility. You may have more information that than the forecaster. Some of the information you use to decide on your tenth bet may arrive after the earlier bets are made and settled. You do not need to specify a whole strategy for betting at the outset, and you do not even need to anticipate what sort of information you might receive. You can decide to stop betting whenever you want (optional stopping). You can decide to continue longer than you might have initially intended (optional continuation).

This fundamental principle is constitutive. It is not derived from some other interpretation of mathematical probability. The very meaning of probabilities in the theory we are teaching is that they will resist testing by betting that follows this fundamental principle.

How strong is the evidence if we multiply our money by two? By five? By ten? Here we need some conventions, informed by experience.

Give artificial numerical examples of successive bets on rain in which the betting score reaches and does not reach a conventional level such as two, five, or ten. Use these to argue for an interpretation of these levels of evidence.

Give similar examples of bets against a financial analyst who predicts earnings announcements.

2.4 The insidious temptation of martingaling

When you run low on capital with successive bets, why not raise more capital (put more money on the table)? Then, at the end when you calculate how much you have multiplied your money, you would use the total capital raised as the denominator. Explain why this does not work, using as examples the classic casino martingale (double your bet until you win) and the d'Alembert. See [7, 8]. Discuss the properties of Borel's martingale [4, 5].

Assignment: select a betting martingale from [9] and find its martingale index.

2.5 Making probability predictions

Discuss different methods at a high level:

1. The physical models used to make weather predictions.
2. Neural networks.
3. Statistical modeling, particularly time-series methods.
4. Defensive forecasting.
5. Averaging predictions: Bayes, betting with expert advice.

Discuss how testing enters into the implementation of each prediction method.

2.6 Testing successive forecasts for a single outcome

We have apps that give a probability for rain tomorrow at 7 am and may change that probability hourly in the course of the day. A rival weather forecaster, say Jane, who thinks she has better probabilities can bet against these changing forecasts just as a trader might bet against the market's changing prices for a stock in the course of a day. Jane buys when she thinks the price is going to go up in the next hour (or minute or millisecond), sells when she thinks it is going to go down. There is a final settlement when everyone sees whether it has rained or not. If Jane avoids martingaling and yet multiplies her initial capital by a substantial factor betting in this way, she has reason to brag that she is a better forecaster than the app.

If a forecaster says 50–50 and sticks with it until the end, Jane will not be able to make money from him until the final settlement. Refusing to change your forecast may make sense in the case of a closely matched ball game or an election with a polarized electorate, at least if forecasting stops as soon as the game begins or the polls open. But when we are predicting rain at 7 am tomorrow, or a champion who will emerge by winning the most games in a series, or a party's nominee for president of the United States who will emerge from a series of primary elections, the forecaster who does not change his probabilities as the outcome comes into view may be vulnerable long before the settlement.

Twenty years ago, the example of changing probabilities might have been seemed both artificial and too complicated for elementary instruction. But today changing probabilities for rain are the most ubiquitous probabilities in many people's lives.

2.7 Testing multiple forecasts of the same outcome

Here begins our consideration of the pitfalls of multiple testing. But the first point to make is the one I made already about testing a single forecaster: proper interpretation of your results depends on what you were trying to accomplish. Reliable understanding by others depends on full and honest disclosure of what you were trying to accomplish.

At one extreme, you have your own probability distribution and use Kelly or half-Kelly or some Kelly cousin your bet against all the forecasters. If you defeat them all, you have some evidence that you are a better forecaster than any of them. If you defeat most of them, you have some evidence that you are among the better forecasters.

At the other extreme, you look for possible (not necessarily likely) weak points in each of the forecasts and bet against them. One forecaster seems to confident, another may be spreading out her probabilities too much. Perhaps you try to arbitrage between a forecaster who gives some event a high probability and another who gives it a low probability. In this case we might say that you are on a fishing expedition, and you should charge the cost of the entire expedition against what you catch. You are testing only the hypothesis that at least one of the forecasters is a bad forecaster.[2] Rather than assessing your success against each forecaster separately, by the factor you multiplied the money you bet against that forecaster, you should assess only your overall success, by the factor by which you multiplied all the capital you invested betting against them.

A Bayesian critic may be tempted to justify the distinction between these two extremes with a just-so story about prior probabilities. But such prior

[2]Can two forecasters who give different forecasters both be good forecasters? Perhaps so in the medium run. Perhaps not in the longer run [21, §10.7].

probabilities have no place in this game-theoretic picture, precisely because they cannot be specified in advance and bet against. Instead, we should consider our different treatments of the two extremes as a matter of first principles.

Statisticians are familiar, of course, with the problem of betting against many forecasters of the same outcome; they call it the problem of testing the goodness-of-fit of a statistical model. Statistical models are usually described as consisting of infinitely many probability distributions. Even if we reduced this to a finite number, could we hope to accomplish much by finding different ways to bet against each one and looking at the ratio by which we multiply our total capital? There seems to be an argument here for finding a single alternative that guides our bet against each one.

2.8 The game-theoretic notion of objective probability

What does it mean for a probability or a probability distribution to be objectively correct? It means that we are predicting that the distribution will withstand any reasonable testing by betting. An objective probability distribution is usually unknown or only partially known. The goal of statistical inference is to learn more about it.

The concept of probability, especially when it is based on betting, is subjective in important respects. There must be a subject, or at least we must imagine a subject, who announces the probabilities and another who challenges them by betting. Their objectivity resides in the inability of the challenger to defeat them and is therefore relative to the information the challenger has and acquires in the course of the betting.

Both the objective probabilities and the information with respect to which they are objective are usually unknown. We can think of the two together as a limit toward which a statistical investigation progresses. This unknown limit is not a single unknown probability distribution relative to an unknown set of information but rather an unknown probability tree — i.e., an unknown stochastic process relative to an unknown filtration (event tree) [16].

This probability tree represents the information and opinions of a real individual nor those of an all-knowing god. Rather, they belong to an imagined being between God and man: a demigod. This picture of objective probabilities as the probabilities of a demigod goes back to Cournot, who called the demigod merely a "superior intelligence" [6, §45].[3]

Discuss role of probabilities that are not close to zero or one.

Discuss the use of causal language when discussing objective probabilities.

2.9 Interval estimation of objective probabilities

Before trying to estimate objective probabilities, statisticians usually adopt a statistical model — a class of probability distributions, say $(P_\theta)_{\theta\in\Theta}$, for an outcome or sequence of outcomes. Adopting the model means assuming that

[3]Some relevant passages are translated from the French in [17].

one of these probability distributions is objectively correct, at least with respect to the information available to us.

Given this assumption, we can obtain an interval estimate for θ by betting against all the probability distributions in the model. The estimate at level $1/\alpha$ consists of those θ for which we do not multiply our capital by $1/\alpha$.

How do we choose the bets? The assumption that one of the distributions is objectively correct seems relevant here; under this assumption it seems that our bet against each P_θ should be guided by the other $P_{\theta'}$. Suggestions that have been put forward include using a maximum likelihood value θ' from a different sample [24] and averaging bets recommended by somewhat similar $P_{\theta'}$ [12]. These methods can be adapted to a sequential setting, where the interval (and usually shrinks) is updated as additional observations are made.

Another method, not adapted to sequential updating, would be to test each θ using nearby θ'. If we are interested in a Gaussian mean, for example, we might use an average of Kelly bets with $\theta \pm \delta$ for some δ of practical significance. This seems especially reasonable if we consider our Gaussian model somewhat conventional rather than exact or nearly so.

Do we have a problem of multiple testing here? By the argument of Lesson 2.7, there is no problem if our bets for testing all the θ are based on a single probability distribution. But if we vary our challenge, testing one θ from the viewpoint of one alternative and another θ from the viewpoint of some other alternative, then we might reasonably be accused of being on a fishing expedition.

But because we are assuming that there is a correct objective θ, we can avoid the real or perceived problem of multiple testing in a way spelled out in [21, §10.3] and [19, §4.2]:

- We imagine that the statistician does not do the testing himself but instructs another player, Skeptic, how to test each θ in the model.
- Some demigod tells Skeptic which θ is correct.
- Skeptic tests only this θ following the statistician's instructions.

The statistician does not see Reality's announcement of the correct θ or Skeptic's bets in response; she sees only the observations. But from this she can calculate how the capital would have been multiplied for each θ, and this allows her to calculate the estimation interval.

Should we call the estimation interval a "confidence interval"? By Markov's inequality, it is a confidence interval according to Jerzy Neyman's definition of the name [14]. On the other hand, we know that the name can be misleading. In [19], I proposed the alternative name "$1/\alpha$ warranty interval", which has the virtue that it demands explanation. The explanation is that if Skeptic gives the statistician the money at the end of time, the statistician has a warranty that her account has been credited with at least $1/\alpha - 1$ if the objectively correct θ is not in the interval. This gloss gives a legitimacy to nested warranty intervals that is not available for nested confidence intervals in Neyman's theory.

2.10 Sampling

Using number theory, mathematicians create experiments that produce numbers with specified objective probabilities. If these numbers are associated with members of a population, this becomes sampling from the population, which leads to warranty intervals for average properties of the population.

2.11 Randomized experimentation

Mention pioneering clinical trials where "choosing every second man" was thought to be sufficient randomization.

2.12 Limiting false discoveries

Report on [23].

2.13 Descriptive probability

Suppose we do not really believe the assumption that one of the P_θ in our model is objectively correct. Suppose the model is merely a convenient way of describing our data — a way of reducing the data, as R. A. Fisher put it in 1922 [10]. Perhaps the model cannot be taken seriously because it describes the data as a random sample, whereas we know it is a convenience sample or an entire population.

Conventional confidence intervals are awkward when we do not believe the model. We are proclaiming confidence in something we do not believe. Betting by testing still makes sense, however, because we can ask which of the P_θ best predicts (i.e., describes) the data. Because the P_θ are in competition with each other, we can have them bet against each other, by Kelly, half-Kelly, or some other Kelly cousin. In [18], I emphasize that when we use Kelly betting we obtain the likelihood intervals that Fisher advocated in 1956 [11].

2.14 Testing multiple forecasters who forecast different outcomes

Here we arrive at the center of the multiple testing problem, already described by Cournot in 1843 [6, §111]. Perhaps you are studying different aspects of the same individuals, perhaps you are performing related experiments, continuing in either case until you obtained a "significant" result. Does testing by betting have anything new to say here?

Averaging betting scores is unlikely to give you the significant results you want, and even this is illegitimate except in the unlikely case where you have planned the sequence of experiments or tests in advance. Otherwise you are martingaling.

Often, it may make the most sense to fall back on Lesson 2.13, testing by betting as descriptive probability. If you have analyzed the entire population, then perhaps you can only leave the matter there. If you can do another experiment or use another body of individuals to replicate your study, then your descriptive analysis can be thought of as an exploratory analysis, with the replication in the confirmatory role.

3 Acknowledgments

This note grew out of conversations at the first workshop on Game-Theoretic Statistics and Safe Anytime Valid Inference (at Eurandom from May 29 to June 3, 2022). It draws particular inspiration from conversations with Wouter Koolen, Volodya Vovk, and particularly Aaditya Ramdas. It has also been influenced by Judith ter Schure, who has emphasized the importance and possibility of teaching game-theoretic statistics at an elementary level. Aaditya, Judith, Volodya, Wouter, and others at the meeting may not, of course, agree with my analyses.

I very much hope for additional suggestions, advice, and other help from colleagues interested in game-theoretic statistics.

References

[1] George Barnard. Review of *An Objective Theory of Probability* by D. A. Gillies. *Journal of the Royal Statistical Society A*, 137(2):262–264, 1974. 3

[2] Jacob Bernoulli. *Ars Conjectandi*. Thurnisius, Basel, 1713. 12

[3] Jacob Bernoulli. *The Art of Conjecturing, together with Letter to a Friend on Sets in Court Tennis*. Johns Hopkins University Press, Baltimore, 2006. Translation of [2] and commentary by Edith Sylla. 3

[4] Émile Borel. Sur une martingale mineure. *Comptes rendus*, 229(23):1181–1183, 1949. 7

[5] Marie-France Bru and Bernard Bru. Émile Borel's denumerable martingales, 1909–1949. In Laurent Mazliak and Glenn Shafer, editors, *The Splendors and Miseries of Martingales: Their History from the Casino to Mathematics*. Birkhäuser, to appear, 2022. 7

[6] Antoine Augustin Cournot. *Exposition de la théorie des chances et des probabilités*. Hachette, Paris, 1843. 9, 11

[7] Harry Crane and Glenn Shafer. Risk is random: The magic of the d'Alembert, 2020. Working Paper #57 at `www.probabilityandfinance.com`. Also at Researchers.One `https://researchers.one/articles/20.08.00007`. 3, 7

[8] Valentin Dimitrov, Glenn Shafer, and Tiangang Zhang. The martingale index, 2022. Working Paper #61 at `www.probabilityandfinance.com`. 3, 7

[9] Stewart Ethier. *The doctrine of chances: Probabilistic aspects of gambling.* Springer, Berlin, 2010. 7

[10] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, 222(602):309–368, 1922. 11

[11] Ronald A. Fisher. *Statistical Methods and Scientific Inference.* Hafner, 1956. Later editions in 1959 and 1973. 11

[12] Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing, 2019. `https://arxiv.org/abs/1906.07801`, conditionally accepted by the *Journal of the Royal Statistical Society B.* 10

[13] John Maynard Keynes. *The General Theory of Employment, Interest and Money.* Macmillan, 1936. 2

[14] Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions Royal Society of London, Series A*, 236:333–380, 1937. 10

[15] Glenn Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, 1976. 2

[16] Glenn Shafer. *The Art of Causal Conjecture.* MIT Press, 1996. 9

[17] Glenn Shafer. Cournot in English, 2018. Working Paper #48 at `www.probabilityandfinance.com`. 9

[18] Glenn Shafer. Descriptive probability, 2021. `http://glennshafer.com/assets/downloads/likelihood.pdf`. 11

[19] Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication (with discussion and response). *Journal of the Royal Statistical Society: Series A*, 184(2):407–478, 2021. 1, 4, 10

[20] Glenn Shafer. Martingales at the casino. In Laurent Mazliak and Glenn Shafer, editors, *The Splendors and Miseries of Martingales: Their History from the Casino to Mathematics.* Birkhäuser, to appear, 2022. 3

[21] Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance.* Wiley, New York, 2019. 8, 10

[22] Judith ter Schure. *ALL-IN meta-analysis.* PhD thesis, University of Leiden, 2022. 1, 2

[23] Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values, 2020. `https://arxiv.org/abs/2009.02824`. 11

[24] Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020. 10