

SIPTA Online School 2020, University of Liverpool Institute for Risk and Uncertainty

Game-theoretic foundations for statistical testing and imprecise probabilities

Remote lectures by **Glenn Shafer**. December 9th and 10th, 2020.

Lecture 1. Testing predictions by betting against them.

Reading: [Testing by betting: A strategy for statistical and scientific communication](#), with discussion and response, by Glenn Shafer. To appear in the *Journal of the Royal Statistical Society, Series A*.

Game-Theoretic Foundations for Probability and Finance

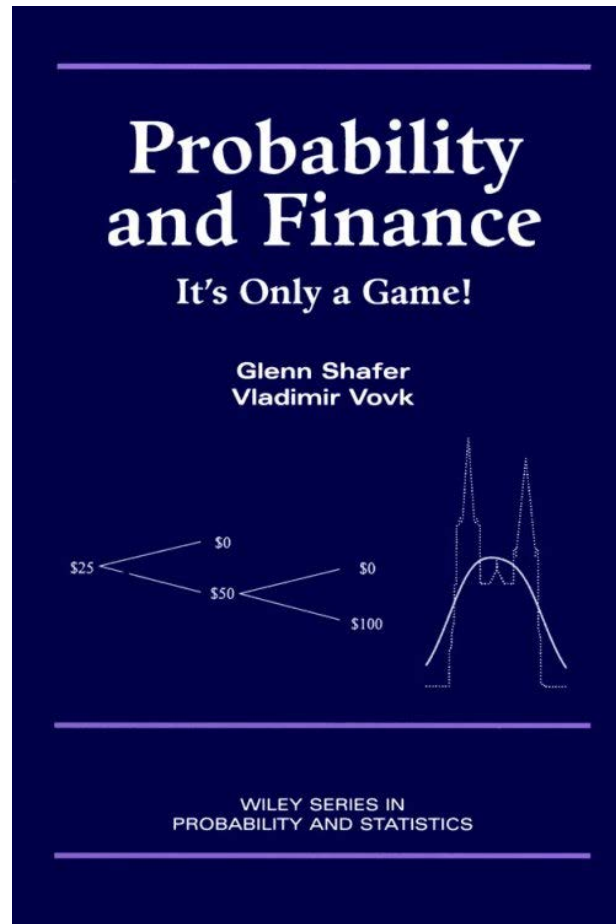
Glenn Shafer | Vladimir Vovk



Bases mathematical probability
on testing by betting.

Working papers at
www.probabilityandfinance.com:

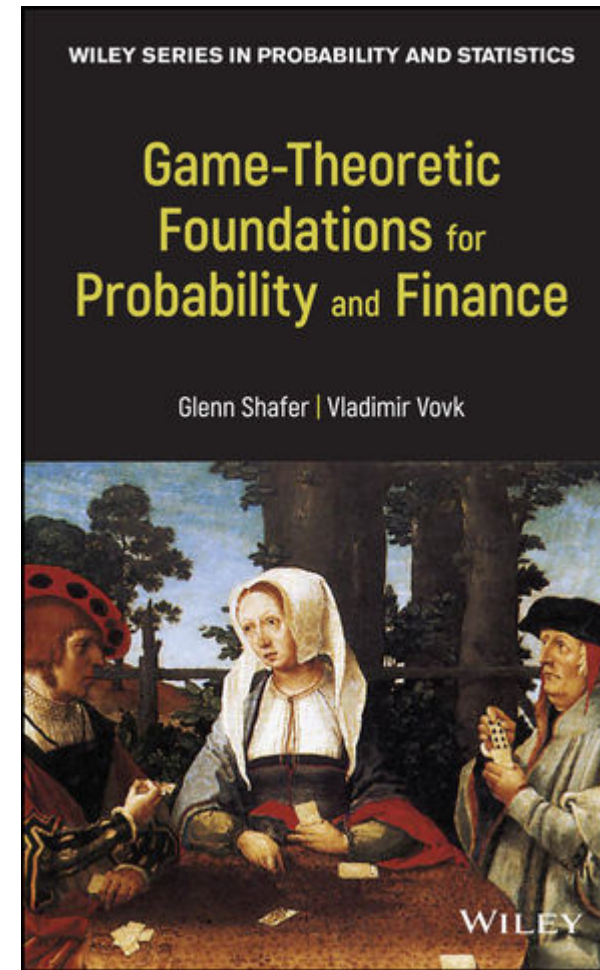
- 47 (efficient markets)
- 54 (testing by betting)
- 55 (history of testing)
- 56 (statistics)
- 57 (random risk)



2001

Showed by example that the classical limit theorems can be proven in game theory.

- Each proof is a betting strategy.
- So more constructive than measure theory.



2019

- Puts game-theoretic probability on a par with measure-theoretic probability as abstract theory.
- New applications (forecasting, decision, CAPM, equity premium, stochastic calculus, calibration, etc.)

1. Testing pundits and weather forecasters

Diversity of probability forecasting

Fundamental principle of testing by betting

Testing by betting vs testing by small probabilities

2. Testing by betting for statisticians

Likelihood ratios

Multiple testing

Replace power with implied target

Three examples

Warranties

3. The deceptiveness of random risk

1. Testing pundits and weather forecasters

The most effective forecast is rarely the one that is connected with high probabilities.

Anders Angström, 1919

The diversity of probability forecasting

Numerical forecasts can be produced by...

- statistical models with estimated parameters
- physical models (hurricane forecasting)
- neural networks
- seat of pants or whatever (financial analyst)

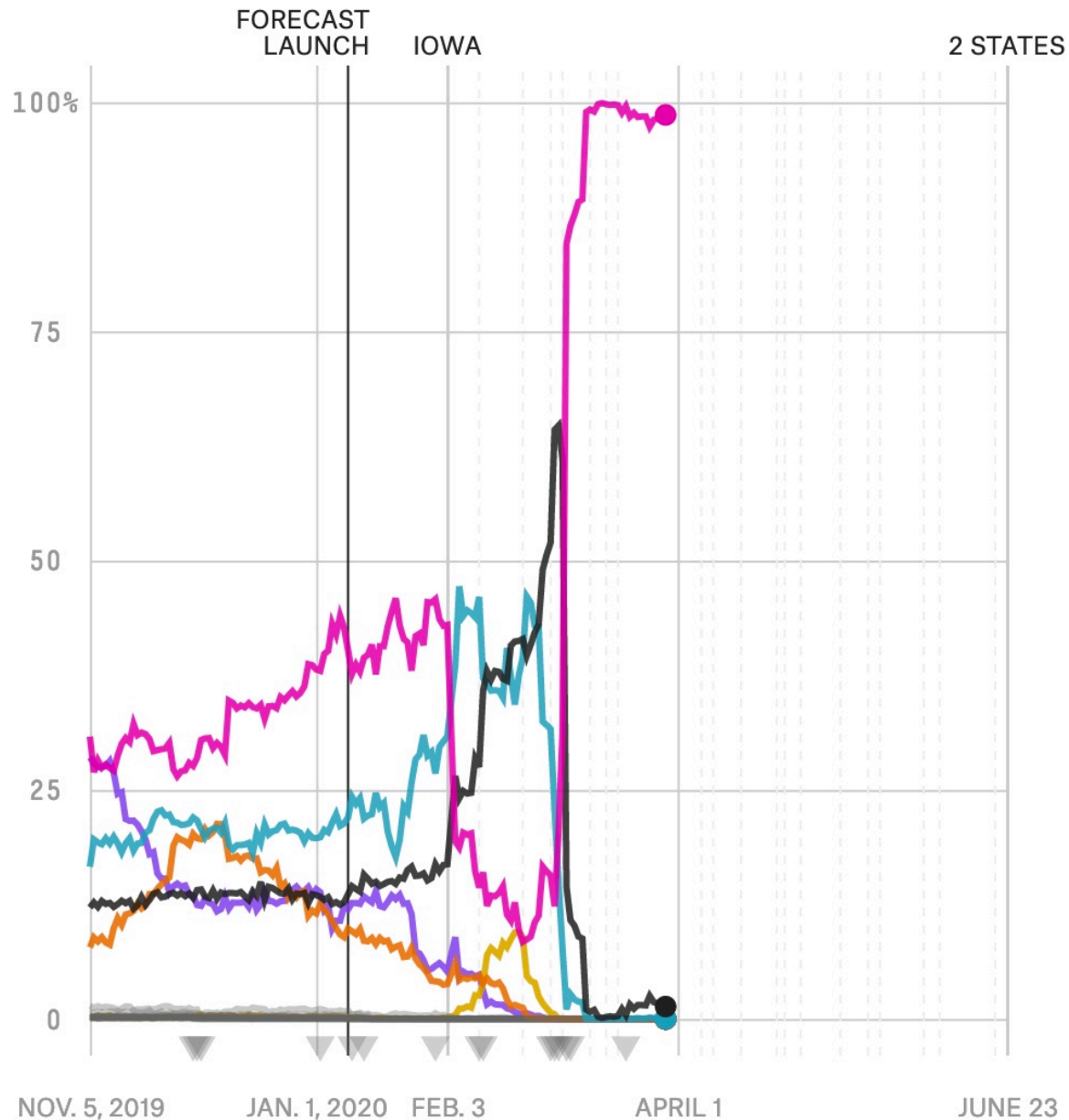
Forecast may be





- a probability (e.g. for rain)
- an estimate (e.g. for size of dividend)

Each forecast may be on a different topic.

We can always test by betting.

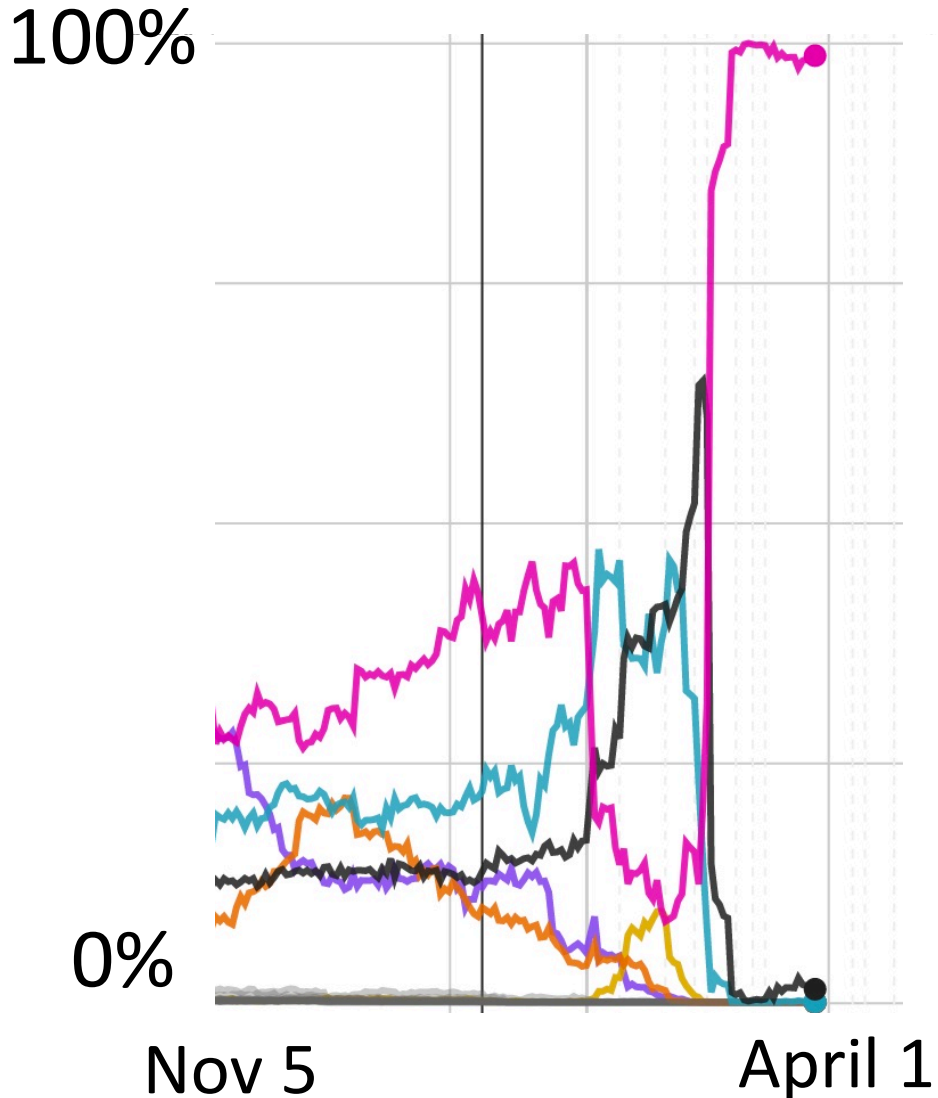
How each candidate's chances of winning more than half of pledged delegates have changed over time



		LATEST ODDS
	Biden	99 in 100 [99%]
	No one	1 in 100 [1%]
	Sanders	<1 in 100 [0.1%]
	Dropouts <input type="checkbox"/>	<1 in 100 [0%]

Screen shot from [fivethirtyeight.com](https://www.fivethirtyeight.com) on March 29

How to test by betting?



Interpret the probabilities as prices.

On December 1,
(probability for Harris is 13%) =
(you can buy or sell **1** if **Harris nominated**, **0** if not for 0.13).

- Each day buy or sell as much of each candidate as you want.
- Next day sell portfolio, buy a new one.
- Start with capital 1.
- Portfolio's maximum net loss exceed current capital.
- Final capital (*betting score*) is evidence against Nate Silver.

Ideas in red new to finance and to statistics.

Fundamental principle of 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.

Fundamental principle of 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.

Starting with unit capital is only for convenience.

Discredit depends on the ratio $(\text{final capital})/(\text{initial capital})$.

Fundamental principle of 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.

If the forecaster keeps forecasting, you can keep betting. Neither of you need to have a plan or strategy about what to forecast, how to forecast, or how to bet.

Fundamental principle of 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.

Each bet uses only the capital remaining from the previous bet. You may not borrow or otherwise raise more capital in order to continue betting.

Fundamental principle of 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**.

You cannot claim full credit for the highest level of capital you reached. You must compare initial with final.

But see WP 34 at www.probabilityandfinance.com.

Fundamental principle of 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.

- If forecaster gives a probability, you can bet on either side at the corresponding odds.
- If forecaster gives a probability distribution, you can buy any payoff for its expected value.
- If forecaster gives an estimate E of an outcome X , you can buy or sell $\$X$ for $\$E$.
- If forecaster gives a new price for A every day, you can buy tomorrow's price for today's.
- If forecaster gives upper and lower previsions, you can buy at the upper.

Fundamental principle of 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.

Not the consequence of some other methodology.

Consistent with “frequentist” practice, but more general.

Another Example: Alice announces probabilities for sports events.

- Week 1: Probabilities of winning for the players in a tennis tournament.
- Week 2: Probabilities for a soccer game: win, lose, or tie.
- Week 3: Probabilities for winning point spread in a cricket game.
- Etc.

How can you test Alice?

You can try to make money at the odds she offers.

Can you think of any other way?

Alice announces probabilities for sports events.

- Week 1: Probabilities of winning for the players in a tennis tournament.
- Week 2: Probabilities for a soccer game: win, lose, or tie.
- Week 3: Probabilities for winning point spread in a cricket game.
- Etc.

Suppose Bob starts with \$100 and does not risk more than that.

- Bob buys age of Wimbledon winner for \$28.

Winner turns out to be 25.

Now Bob has \$97.

- Bob pays \$97 for (\$0 if Madrid, \$100 if Barcelona or tie).

Madrid wins.

Now Bob has \$0.

Now Bob has to stop betting, because he is out money.

Bob is not allowed to risk more than his original \$100.

Bob can challenge and discredit Alice without giving alternative probabilities.

Maybe he does not believe that there are meaningful or reliable probabilities for the events in question.

Bob can bet with play money.

His goal is to make a point, not to get rich.

No real counterparty to his play-money bets.

Alice is not risking real money either; she is risking only her reputation as a forecaster.

People understand the significance of such betting outcomes.

1. Alice may know more than Bob. If Bob makes money, then perhaps Alice's additional information is not worth much.
2. Bob may know more than Alice. If Bob makes money on her forecasts, then his extra information may be relevant.
3. If Bob does not make money, then we have no evidence against Alice's probabilities. If Bob is clever and knowledgeable, then we even have evidence in Alice's favor.

You can test by betting even when Alice does not give a full probability distribution.

- Alice's earnings forecast is the price of the actual earnings number.
- Today's stock price is the price of tomorrow's stock price.

In *Game-Theoretic Foundations for Probability and Finance*, we

- test market efficiency by betting,
- use resistance to such testing as a definition of market efficiency,
- derive properties of market prices (equity premium, fluctuation, etc.)

What principle should be used to translate mathematical probability theorems into real life? In 1976, the probabilist J. L. Doob asked this question and answered by explaining that the statistician...

sets up a model and comes to operational decisions based on the principle that **hypotheses must be reexamined if they ascribe small probability to a key event that actually happens.**

Doob's principle has many names. For now, let's call it the principle of testing-with-small-probabilities.

How is the **principle of testing-with-small-probabilities** related to the **principle of testing-by-betting**?

Markov's inequality is part of the answer.

Markov's inequality says that when S is a non-negative payoff with expected value 1, $\mathbf{P}(S \geq c) \leq 1/c$.

You can sometimes use Markov's inequality to justify testing by betting:

There is at best one chance in 100 that I can multiply my money by 100.

But this applies only when the predictions were given by a global \mathbf{P} .

Game-theoretic probability turns Markov on his head. Instead of taking **happening of event of small purported probability** as evidence against predictions,

- we take **multiplying money risked by a large factor** as evidence against predictions.
- we use betting to **define** global probabilities from the predictions.

Two flavors of principle of testing-with-small-probabilities:

Fixed-level test. Fix key event to which forecaster gives small probability α . Say forecaster is discredited at level α if E happens.

p-value. Fix test statistic T that forecaster says should not be too large. Observe value t . Consider smallness of forecaster's $\mathbf{P}(T \leq t)$ measure of evidence against forecaster.

Intuitively, fixed-level discrediting at level α is stronger evidence than an equally small p-value.

A fixed level test can be thought of as a bet at odds $\alpha/(1 - \alpha)$. (More explanation later.) The notion of a p-value can be related to testing-by-betting via Ville's inequality, which generalizes Markov's inequality.

Markov's inequality. If S is a nonnegative random variable and $E_P(S) = 1$, then

$$P(S \geq c) \leq \frac{1}{c}.$$

The event $S \geq \alpha$ can be used as a fixed level α test.

Ville's inequality. Suppose Y_1, Y_2, \dots is a stochastic process, and you bet on the Y_n in order, starting with capital 1 and following a strategy that always keeps your capital nonnegative no matter how the bets come out. Let S_1, S_2, \dots be the resulting capital process (nonnegative martingale). Then

$$P(S_n \geq c \text{ for some } n) \leq \frac{1}{c}.$$

When $\max_n S_n$ is the test statistic, the p-value is $1 / \max_n S_n$.

2. Testing by betting for statisticians

Being, therefore, unable to get a mathematical measure of the assurance with which we may accept our estimate because we do not first possess a mathematical measure of its inherent plausibility, we turn to the task of finding the best possible makeshift. Fortunately, the makeshift is not a very unsatisfactory one when its meaning is clearly understood...

Thomas C. Fry, 1928

Hypothesis: P describes random variable Y .

Question: How do we use $Y = y$ to test P ?

Conventional answer:

- Choose *significance level* α , say 0.05.
- Choose E such that $P(E) = 0.05$.
- Reject P if $y \in E$.

Hypothesis: P describes random variable Y .

Question: How do we use $Y = y$ to test P ?

Conventional answer:

- Choose *significance level* α , say 0.05.
- Choose E such that $P(E) = 0.05$.
- Reject P if $y \in E$.

Betting interpretation:

- Put £1 on E .
- Get back £0 if E fails.
- Get back £20 if E happens.
 - You multiplied your money by a large factor.
 - This discredits P .
 - What better evidence could you have?

Question: How do we measure the strength of evidence against P ?

Conventional answer:

- Use a test statistic to define a test for each $\alpha \in (0, 1)$.
- The *p-value* is the smallest α for which the test rejects.
- The smaller the p-value, the more evidence against P .

Too complicated!

Question: How do we measure the strength of evidence against P ?

Conventional answer:

- Use a test statistic to define a test for each $\alpha \in (0, 1)$.
- The *p-value* is the smallest α for which the test rejects.
- The smaller the p-value, the more evidence against P .

Betting alternative:

Make a bet on Y that can pay many different amounts

- Such a bet is a function $S(Y)$.
- Choose S so that $E_P(S) = 1$.
- Pay $\pounds 1$ and get back $\pounds S(y)$.
- The larger $S(y)$, the more evidence against P .

Call $S(y)$ the *betting score*.

This is the factor by which you multiplied your money.

If $E_P(S) \neq 1$, betting score is

$$\frac{S(y)}{E_P(S)}.$$

Likelihood ratios

A **betting score**, as just defined, is the same thing as a likelihood ratio.

- A **bet** S is a function of Y satisfying $S \geq 0$ and $\sum_y S(y)P(y) = 1$.
- So SP is also a probability distribution. Call it the **alternative** Q .
- But $Q(y) = S(y)P(y)$ implies $S(y) = Q(y)/P(y)$.
- A bet against P defines an alternative Q and the betting score $S(y)$ is the likelihood ratio $Q(y)/P(y)$.

Conversely, if you start with an alternative Q , then Q/P is a bet.

Proof:

$$\frac{Q(y)}{P(y)} \geq 0 \text{ for all } y.$$

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

But is wanting to test against Q good reason for using the bet Q/P ?

Multiple testing

Fundamental principle of 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.

Suppose you use Y_1 to test P , obtaining a betting score $S_1(y_1)$. Finding this result inconclusive, you observe Y_2 and use it to test P again, this time obtaining a betting score $S_2(y_2)$.

Your initial capital of 1 became $S_1(y_1)$ after the first bet. This is all you have to invest in the second bet, and the meaning of the second score is that you multiply this investment by $S_2(y_2)$.

So your score is now $S_1(y_1)S_2(y_2)$.

You say P describes Y .

I want to bet against you.

I think Q describes Y .

Should I use Q/P as my bet?

$S = Q/P$ maximizes $\mathbf{E}_Q(\ln S)$.

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

Gibbs's inequality

Why maximize $\mathbf{E}_Q(\ln S)$? Why not $\mathbf{E}_Q(S)$? Or $Q(S \geq 20)$?

Neyman-Pearson lemma

When S is the product of successive factors, $\mathbf{E}(\ln S)$ measures the rate of growth (Kelly, 1956). This has been used in gambling theory, information theory, finance theory, and machine learning. Here it opens the way to a theory of multiple testing and meta-analysis.

Replace power with *implied target*.

The *implied target* of the test $S = Q/P$ is $\exp(E_Q(\ln S))$.

$$\mathbf{E}_Q(\ln S) = \sum_y Q(y) \ln S(y) = \sum_y P(y) S(y) \ln S(y) = \mathbf{E}_P(S \ln S)$$

Use the implied target to evaluate the test in advance.

Even if I do not take Q seriously, my critics will.

Why should the editor invest in my test if it is unlikely to produce a high betting score even when it is optimal?

Elements of a study that tests a probability distribution by betting

	name	notation
Proposed study		
initially unknown outcome	phenomenon	Y
probability distribution for Y	null hypothesis	P
nonnegative function of Y with expected value 1 under P	bet	S
$S \times P$	implied alternative	Q
$\exp(\mathbf{E}_Q(\ln S))$	implied target	S^*
Results		
actual value of Y	outcome	y
factor by which money risked has been multiplied	betting score	$S(y)$

Three examples

In the following examples,

- we have only a single observation from a normal distribution, and
- the null and alternative differ only in their mean.

So are they merely “toy examples”?

No! They are very, very general.

1. The most widely used statistical test is for the difference between two proportions (% cured by treatment - % cured by placebo, for example). This test uses the normal approximation, and the null hypothesis is that the resulting normal observation has mean zero.
2. You get the same picture when the test statistic is the average of n observations.
3. The issues illustrated can arise in almost any statistical test.

Example 1.

Result statistically and practically significant but hopelessly contaminated with noise.

$$P: Y \sim \mathcal{N}(0, 10)$$

$$Q: Y \sim \mathcal{N}(1, 10)$$

$$y = 30$$

$P: Y \sim \mathcal{N}(0, 10)$

$Q: Y \sim \mathcal{N}(1, 10)$

$$y = 30$$

- p-value: $P(Y \geq 30) \approx 0.00135$.
- 5% test rejects when $y \geq 16.445$.
Power 6%.
- Bet Q/P has implied target 1.005.
Betting score is $S(30) \approx 1.34$.

- Power and implied target agree: study is worthless.
- But Neyman-Pearson rejects with low p-value, while betting score sees that evidence is slight.

Example 2.

Test with $\alpha = 5\%$ and high power rejects with borderline outcome even though likelihood ratio favors null.

$$P: Y \sim \mathcal{N}(0, 10)$$

$$Q: Y \sim \mathcal{N}(37, 10)$$

$$y = 16.5$$

- p-value: $P(Y \geq 16.5) \approx 0.0495$.
- 5% test rejects when $y \geq 16.445$.
Power 98%.
- Bet Q/P has implied target 939.
Betting score is $S(16.5) \approx 0.477$.

$$P: Y \sim \mathcal{N}(0, 10)$$

$$Q: Y \sim \mathcal{N}(37, 10)$$

$$y = 16.5$$

- Power and implied target agree: study is good.
- Neyman-Pearson rejects.
Betting score says evidence slightly favors null.

Example 3.

High p-value is interpreted as evidence for null.

$$P: Y \sim \mathcal{N}(0, 10)$$

$$Q: Y \sim \mathcal{N}(20, 10)$$

$$y = 5$$

$P: Y \sim \mathcal{N}(0, 10)$

$Q: Y \sim \mathcal{N}(20, 10)$

$$y = 5$$

- p-value: $P(Y \geq 5) \approx 0.3085$.
- 5% test rejects when $y \geq 16.445$.
Power 64%.
- Bet Q/P has implied target 7.39.
Betting score is $S(5) \approx 0.368$.

- Power and implied target agree: study is marginal.
- Neyman-Pearson simply does not reject.
Betting score says evidence slightly favors null.

Warranties

Statistical work often begins with an indexed class of probability distributions on a space \mathcal{Y} , say

$$(P_\theta)_{\theta \in \Theta}.$$

This a *statistical model* or a *parametric model*, θ being the parameter.

Statistical model $(P_\theta)_{\theta \in \Theta}$

- A statistician can use observations to test each P_θ at the same level α .
- The subset of $A \subseteq \Theta$ consisting of θ for which P_θ passes is called a $(1 - \alpha)$ -*confidence set*.

The subset of $A \subseteq \Theta$ consisting of θ for which P_θ passes is called a $(1 - \alpha)$ -*confidence set*.

Interpreting each test in betting terms, we can also say that A is the set of θ for which the bettor did not multiply her money by $1/\alpha$. With this interpretation, we call A a $(1/\alpha)$ -*warranty set*.

In the context of a single set of observations on a parametric model, a $(1/\alpha)$ warranty set is the same thing as a $(1 - \alpha)$ -confidence set.

But when we combine different studies, the testing-by-betting principle authorizes us to multiply the betting scores for each θ , so we obtain $(1/\alpha)$ warranty sets that are not $(1 - \alpha)$ -confidence sets.

3. The deceptiveness of random risk

Donnez moi, disoit Monsieur P... à Mr. Huyn, donnez moi un jeu dont les diverses chances finissent, de tems en tems par s'égaliser, je trouverai une manière régulière, simple et facile d'y jouer avec assurance de gain: au moyen de l'application d'une Martingale graduée.

James Smyll, 1820

Popular betting games in casinos of yesteryear

- 18th century: Trente et Quarante (French for 30 and 40)
- 19th century: Roulette

Trente et Quarante

Casino's advantage $\approx 1\%$

Simplest bets are even-money bets on **red or black**.

T is the Tailleur, who dealt the cards.

C is the Croupier, who moved the money from losers to winners.



Roulette

You can still bet on **red or black**.

Casino's advantage = $2/38 \approx 5\%$

No one would bet on a coin flip in a casino.



In a standard deck of playing cards (invented by the French in the 1400s), half the cards are red and half are black.

Between friends, you might make an even-money bet on red or black by drawing a card at random.

But no one would trust the casino's *tailleur* to draw a card at random!

Trente et Quarante

Deal two rows of cards.

Call one row **red**, the other **black**.

In each row, stop dealing when the sum > 30 .

Ace is 1; face card is 10.

The row whose total is closest to 30 wins!

Ignore ties.

But when the tie is 31-31, the casino gets half the money.

This is the 1% advantage.

Table for Trente et Quarante

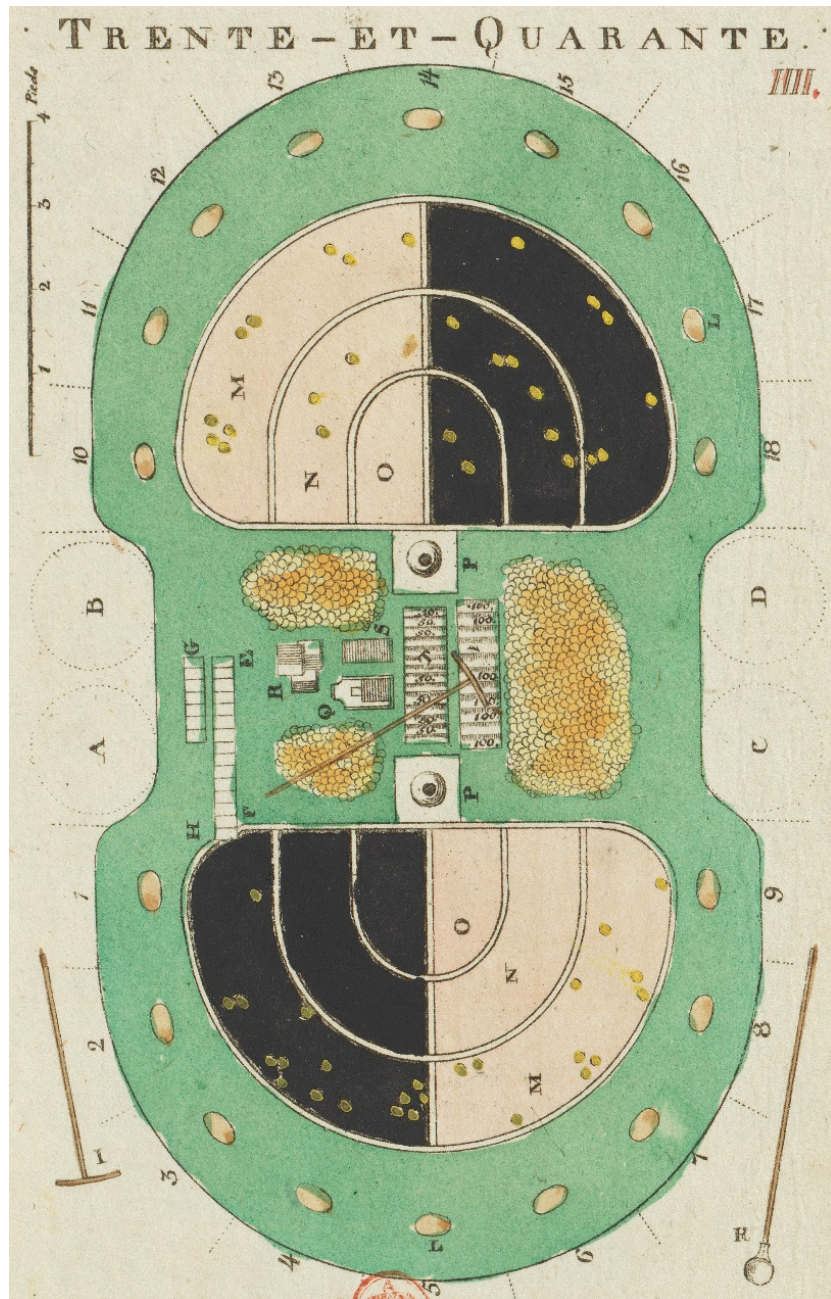
To bet, put your money on the table.

Keep your money (including winnings) on the green.

To bet, push money into the closest black or red box.

When behind, take more money out of your pocket to keep betting.





In elegant legal casino

The seated players keep their winnings in the yellow cells on the green.

How large do you expect R to be in the casino?

You invest K . You make a net profit of G .

Three ways of measuring your success:

- Return: $R = \frac{G}{K}$
- Logarithmic return: $\ln(1 + R)$
- Score: $1 + R = \frac{K + G}{K}$

The casino has an advantage.

$$\mathbf{E}(G) < 0.$$

So $\mathbf{E}(R)$ is also negative:

$$\mathbf{E}(R) = \mathbf{E}\left(\frac{G}{K}\right) = \frac{\mathbf{E}(G)}{K} < 0.$$

How large do you expect R to be in the casino?

You invest K . You make a net profit of G .

Three ways of measuring your success:

- Return: $R = \frac{G}{K}$
- Logarithmic return: $\ln(1 + R)$
- Score: $1 + R = \frac{K + G}{K}$

MARKOV'S INEQUALITY

If bets are fair,

$$\mathbf{P}(1 + R \geq c) \leq \frac{1}{c}$$

for all $c > 0$.

This also true, of course, when the casino has an advantage.

You invest K . You make a net profit of G .

Three ways of measuring your success:

- Return: $R = \frac{G}{K}$
- Logarithmic return: $\ln(1 + R)$
- Score: $1 + R = \frac{K + G}{K}$

The casino has an advantage.

$$\mathbf{E}(G) < 0.$$

So $\mathbf{E}(R)$ is also negative:

$$\mathbf{E}(R) = \mathbf{E}\left(\frac{G}{K}\right) = \frac{\mathbf{E}(G)}{K} < 0.$$

$$\mathbf{P}(1 + R \geq c) \leq \frac{1}{c} \text{ for all } c > 0.$$

Here's the catch:

These results depend on K being a constant.

Is K constant in the casino?



To bet, put your money on the table.

Keep your money (including winnings) on the green.

To bet, push money into the closest black or red box.

When behind, take more money out of your pocket to keep betting.

When behind take more money out of your pocket to keep betting.

You never told anyone how much you had in your pocket or how much you are willing to risk.

Maybe you don't know yourself.

You use the amount you actually take out of your pocket as K .
But this is random!

Because you put more money on the table when you are behind, K is negatively correlated with G . This leads to

$$\mathbf{E}(R) > 0$$

in spite of the house's advantage. If you can keep putting money on the table until you quite when you are ahead, you also get

$$\mathbf{P}(R > 0) \text{ near } 1.$$

SIMPLE EXAMPLE

- I make a \$1 bet 10 times.
- First I put \$1 on the table for the first bet.
- I put more on the table only when needed.

A few insights:

- The more I lose (smaller G), the more I put on the table (larger K).
- If $K > 5$, then $G < 0$.
- If I have early wins, I can weather later losses without more out of pocket. If I take only 1 out of pocket and net 2, $G/K = 2$.
- I can't lose more than I take out of my pocket. The worst possible G/K is -1 .

The **d'Alembert** was the most popular 19th-century betting system.

Start by betting 1 unit.

- When you lose, increase your bet by 1 unit.
- When you win, decrease your bet by 1 unit, unless it is already only 1 unit.
- Stop when you are 4 units ahead or after 50 bets, whichever comes first.

This has expected return over 100% and a 98% chance of winning something. These numbers do not change much when the house has a 2% to 4% advantage, as in Roulette.

**When K is also random
and not independent of G .**

Suppose $\mathbf{Cov} \left(G, \frac{1}{K} \right) > 0.$ (*)

Then the expected return is positive.

In fact, (*) is the expected return:

$$\mathbf{Cov} \left(G, \frac{1}{K} \right) = \mathbf{E} \left(\frac{G}{K} \right) - \mathbf{E}(G)\mathbf{E} \left(\frac{1}{K} \right).$$

Where do we see enterprises raising more money when they are behind?

- Start-ups
- Hedge funds with huge losses.
 - Long-Term Capital Management (1998)
 - MF Global (Jon Corzine, 2011)
- Investments by corporations
- Investments by governments
- Both mutual fund reports and academic studies most often use a definition of “return” that does not put all the money risked in the denominator. (Options, short-selling.)