Testing by betting in statistical inference

Glenn Shafer
April 18, 2025

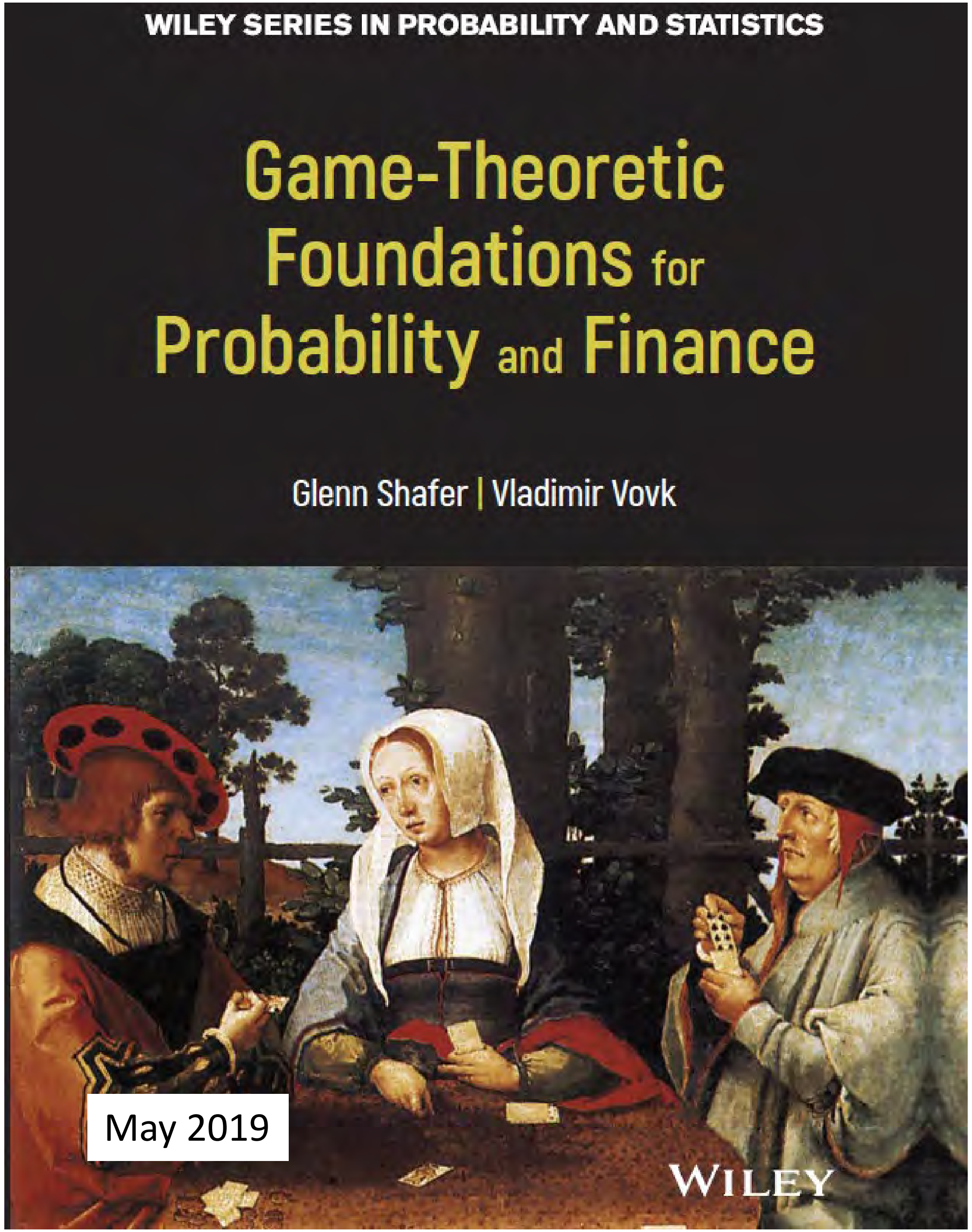
University of Kansas, School of Business

Test forecaster (or forecasting strategy) by betting against it.

If it passes, use its forecasts (or strategies) to make predictions.

Do statistical inference this way.

My forthcoming book on game-theoretic statistics.



Bases **mathematical probability** on testing by betting.

Working papers at
www.probabilityandfinance.com

How to apply the ideas to statistics?

Two approaches

Conventional

Use testing by betting to define Neyman-Pearson tests.

- Brief elementary summary: https://probabilityandfinance.com/misc/snapshot.pdf.
- Book, mathematically advanced: *Testing Hypotheses with E-values*, by Aaditya Ramdas and Ruodu Wang, https://arXiv.org/abs/2410.23614

Game-theoretic

Use betting game directly.

- Testing by betting: a strategy for statistical and scientific communication, with discussion and response. Journal of the Royal Statistics Society, Series A 184(2):407-478, 2021.
- Book, more elementary: *An Introduction to Game-Theoretic Statistics (under preparation)*, by Glenn Shafer

The game

Initial capital

PLAYERS: Skeptic, Forecaster, Reality

PARAMETER: Probability space $\mathcal{Y}$

Skeptic announces $s_0 \in \mathbb{R}$.

Forecaster announces probability distribution P on $\mathcal{Y}$.

Skeptic announces variable Z on $\mathcal{Y}$ with finite $\mathbf{E}_P(Z)$.

Reality announces $y \in \mathcal{Y}$.

$s := s_0 + Z(y) - \mathbf{E}_P(Z)$.

Final capital

Skeptic bets with play money.

Betting score = s/s_0

PLAYERS: Skeptic, Forecaster, Reality
PARAMETER: Probability space $\mathcal{Y}$

Skeptic announces $s_0 \in \mathbb{R}$.
Forecaster announces probability distribution P on $\mathcal{Y}$.
Skeptic announces variable Z on $\mathcal{Y}$ with finite $\mathbf{E}_P(Z)$.
Reality announces $y \in \mathcal{Y}$.
$s := s_0 + Z(y) - \mathbf{E}_P(Z)$.

The game is in Statistician's imagination.
She tells Skeptic how to bet.
Sometimes she tells Forecaster how to forecast.

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.

An Introduction to Game-Theoretic Statistics

This is a work in progress. Chapters 8 and 9 are especially incomplete.
I welcome comments and suggestions for improvement.

The statistician can work inside the game.

PLAYERS: Skeptic, Experimenter, Forecaster, Reality

Skeptic announces $s_0 \in \mathbb{R}$.

For $n = 1, 2, \ldots$:

Experimenter announces probability space $\mathcal{Y}_n$.

Forecaster announces probability distribution P_n on $\mathcal{Y}_n$.

Skeptic announces variable Z_n on $\mathcal{Y}_n$ with finite $\mathbf{E}_{P_n}(Z_n)$.

Reality announces $y_n \in \mathcal{Y}_n$.

$s_n := s_{n-1} + Z_n(y_n) - \mathbf{E}_{P_n}(Z_n)$.

Regression

PLAYERS: Skeptic, Informant, Reality, Forecaster
PARAMETERS: Set $\mathcal{X}$, probability space $\mathcal{Y}$

Skeptic announces $s_0 \in \mathbb{R}$.
Informant announces $x \in \mathcal{X}$.
Forecaster announces probability distribution P on $\mathcal{Y}$.
Skeptic announces variable Z on $\mathcal{Y}$ with finite $\mathbf{E}_P(Z)$.
Reality announces $y \in \mathcal{Y}$.
$s := s_0 + Z(y) - \mathbf{E}_P(Z)$.

To test, Skeptic must keep his capital non-negative.

PLAYERS: Skeptic, Forecaster, Reality
PARAMETERS: $N \in \mathbb{N}$, probability space $\mathcal{Y}$

$s_0 := 1$.
For $n = 1, \ldots, N$:
Forecaster announces probability distribution P_n on $\mathcal{Y}$.
Skeptic announces nonnegative variable Z_n on $\mathcal{Y}$
such that $\mathbf{E}_{P_n}(Z_n) = s_{n-1}$.
Reality announces $y_n \in \mathcal{Y}$.
$s_n := Z_n(y_n)$.

Betting score

Kelly testing. Forecaster announces probability distribution P. Skeptic thinks R is a better forecast and announces $Z = R/P$, which is permitted is because $Z \geq 0$ and

$$\mathbf{E}_P(Z) = \sum_y \frac{R(y)}{P(y)} P(y) = \sum_y R(y) = 1.$$

Example. Forecasters say probability of rain is $1/2$. Skeptic says $2/3$. So Skeptic pays 1 and gets back.

$$Z(y) = \begin{cases} \frac{2/3}{1/2} = \frac{4}{3} & \text{if } y = \text{rain} \\ \frac{1/3}{1/2} = \frac{2}{3} & \text{if } y = \text{no rain} \end{cases}$$

A betting score against a probability distribution is 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 $Q(y)/P(y)$.

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.

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
$\exp\left(\mathbf{E}_Q(\ln S)\right)$	implied target	S^*
Results		
actual value of Y	outcome	y
factor by which money risked has been multiplied	betting score	$S(y)$

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 (final capital)/(initial capital).

Fundamental principle of testing-by-betting

<u>Successive bets</u> 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.

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.

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

- Interpret an earnings forecast as 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.)

PLAYERS: Skeptic, Forecaster, Reality

Skeptic announces $s_0 \in \mathbb{R}$.
Forecaster announces $\mu \in [-1, 1]$.
Skeptic announces $z \in \mathbb{R}$.
Reality announces $y \in [-1, 1]$.
$s := s_0 + z(y - \mu)$.

PLAYERS: Oracle, Forecaster, Skeptic, Reality

$s_0 = 1$.
Oracle announces $\mu \in \mathbb{R}$.
Forecaster announces probability distribution P on $\mathbb{R}$.
Skeptic announces nonnegative variable Z with $\mathbf{E}_P(Z) = 1$.
Reality announces $y \in \mathbb{R}$.
$s := Z(y)$.

PLAYERS: Oracle, Skeptic, Reality

$s_0 = 1$.
Oracle announces $\mu \in \mathbb{R}$.
Skeptic announces nonnegative variable Z on $\mathbb{R}$ with $\mathbf{E}_{\boldsymbol{N}_{\mu,1}}(Z) = 1$.
Reality announces $y \in \mathbb{R}$.
$s := Z(y)$.

PLAYERS: Oracle, Skeptic, Reality

$s_0 = 1$.

Oracle announces $\mu \in \mathbb{R}$.

Skeptic announces nonnegative variable Z on $\mathbb{R}$ with $\mathbf{E}_{\boldsymbol{N}_{\mu,1}}(Z) = 1$.

Reality announces $y \in \mathbb{R}$.

$s := Z(y)$.

Statistician also assigns Skeptic a strategy. A strategy for Skeptic in Protocol 9.2 is a mapping $\mathcal{Z}$ that assigns to each $\mu' \in \mathbb{R}$ a nonnegative variable $\mathcal{Z}_{\mu'}$ satisfying $\mathbf{E}_{\boldsymbol{N}_{\mu',1}}(\mathcal{Z}_{\mu'}) = 1$.

Warrantied interval: $$W_K := \{\mu' \in \mathbb{R} \mid \mathcal{Z}_{\mu'}(y) < K\}.$$

$$W_K := \{\mu' \in \mathbb{R} \mid \mathcal{Z}_{\mu'}(y) < K\}.$$

How we can use subjective distributions and Bayes's rule.

$$\mathcal{Z}_\mu := \frac{\boldsymbol{n}_{\lambda,\tau^2+1}}{\boldsymbol{n}_{\mu,1}}$$

$$W_K := \left\{\mu' \in \mathbb{R} \,\middle|\, (y-\mu')^2 < \ln(K^2(\tau^2+1)) + \frac{(y-\lambda)^2}{\tau^2+1}\right\}.$$

Too many statistical studies pretend to make inferences from a random sample to a larger population when either

1. the sample is not random (convenience sample) or

2. there is no larger population.

A fictional study of perceptions

Table 1: 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.

	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	$\frac{28}{60} \approx 47\%$	$\frac{32}{140} \approx 23\%$	$\frac{60}{200} = 30\%$

Assume 100% response; this is entire population of 200 employees. Randomness assumption untenable.

	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	$\frac{28}{60} \approx 47\%$	$\frac{32}{140} \approx 23\%$	$\frac{60}{200} = 30\%$

Row (male vs female) and column (BIPOC vs White) differences both significant.

20%-point difference between BIPOC males and BIPOC females has standard error

$$\sqrt{\frac{12}{33}\left(\frac{1}{20}+\frac{1}{10}\right)} \approx 18\ \%\ \text{points.}$$

Basic idea of game-theoretic descriptive statistics

Adopt a statistical model as a convention, not as a hypothesis about a larger population.

Have the different parameter values bet against each other on the data we want to describe.

Give an interval of values that performed best.

Using cutoffs suggested by Fisher in 1956, we may classify the forecasters according to the ratio of their likelihood to that of the winner:

Relatively good. Those who did at least half as well.

Relatively fair. Less than half but at least (1/5)th as well.

Relatively poor. Less than (1/5)th but at least (1/15)th as well.

Unacceptable. Worse than (1/15)th as well.

These cutoffs are arbitrary, but no more so than the 5% and 1% frequencies used for statistical significance. If equally accepted as conventions, they can be equally serviceable. Their meaning in terms of betting will be readily understood by the public.

	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	$\frac{28}{60} \approx 47\%$	$\frac{32}{140} \approx 23\%$	$\frac{60}{200} = 30\%$

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 $\theta = (\theta_{\text{bf}}, \theta_{\text{bm}}, \theta_{\text{wf}}, \theta_{\text{wm}})$ 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 $\theta_{\text{bf}} - \theta_{\text{bm}}$. 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.

Martingaling

JDM (2025), **xx**:xx 1–23

CAMBRIDGE
UNIVERSITY PRESS

RESEARCH PAPER

The Martingale Index: A Measure of Self-Deception in Betting and Finance

Valentin Dimitrov[1] and Glenn Shafer[2]

[1]Rutgers Business School, Newark, 07102, New Jersey, USA. E-mail: vdimitr@rutgers.edu.
[2]Rutgers Business School, Newark, 07102, New Jersey, USA. E-mail: gshafer@rutgers.edu.

Keywords: expected return, martingale index, investment strategies, betting

To appear in Judgement & Decision Making

Abstract
People who repeatedly risk money, whether they be traders for financial institutions, corporate executives, day traders, or sports bettors, sometimes appear to do better than chance only because the risk of large losses is hidden or overlooked. As students of casino gambling know, one way to obscure the risk of large losses is to bet more when you are losing and less when you are winning. In 19th century casinos, betting strategies that did this were called *martingales*. Following such strategies, whether deliberately or unwittingly, was called *martingaling*. Traders in financial instruments often martingale; in fact, they are martingaling whenever they respond to a margin call. A businessperson who doubles down on an apparently losing investment is martingaling. Opinionated sports bettors easily fall into martingaling. The *martingale index*, defined in this paper, measures the portion of the apparent success of a betting, trading, or investment strategy that can be attributed to martingaling. We calculate the martingale index for some popular casino strategies and also for some strategies that model random trading in S&P 500 futures and in stocks. And we discuss how educating the public about the martingale index might help both businesses and individuals avoid the temptations of martingaling.

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 ≈ 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 ≈ 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.

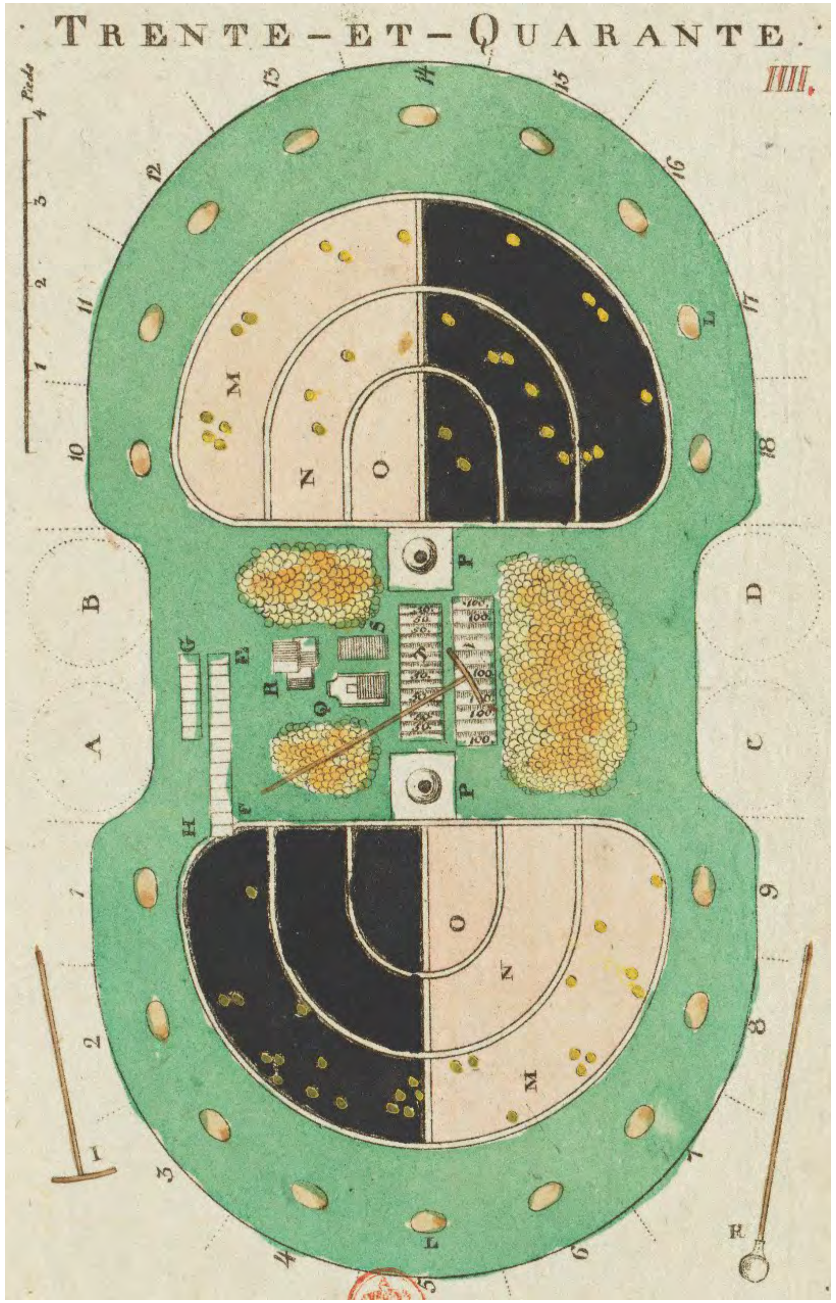
Table for Trente et Quarante

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

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

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:

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

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

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

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.)

Extra slides

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?

Bob tests Alice by betting.

- Starts with $100.

- Buys age of Wimbledon winner for $28.
  Winner turns out to be 25.
  Now Bob has $97.

- Pays $97 for ($0 if Madrid, $100 if Barcelona or tie).
  Madrid wins.
  Now Bob has $0.

- Stops, because he is out money.

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

Bob bets with play money.

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

Alice 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.

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 £1 and get back £$S(y)$.
- The larger $S(y)$, the more evidence against P.

Three examples of hypothesis testing

Example 1.

Result statistically and practically significant but hopelessly contaminated with noise.

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

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

$$y = 30$$

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 α = 5% and high power rejects with borderline outcome even though likelihood ratio favors null.

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

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

$$y = 16.5$$

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 $S(16.5) \approx 0.477$.

- 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.

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

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

$$y = 5$$

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

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

$$y = 5$$

- 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.