Testing by Betting

Glenn Shafer Rutgers University

Royal Statistical Society Discussion Meeting 2 September 2020

How can you test probabilistic predictions?

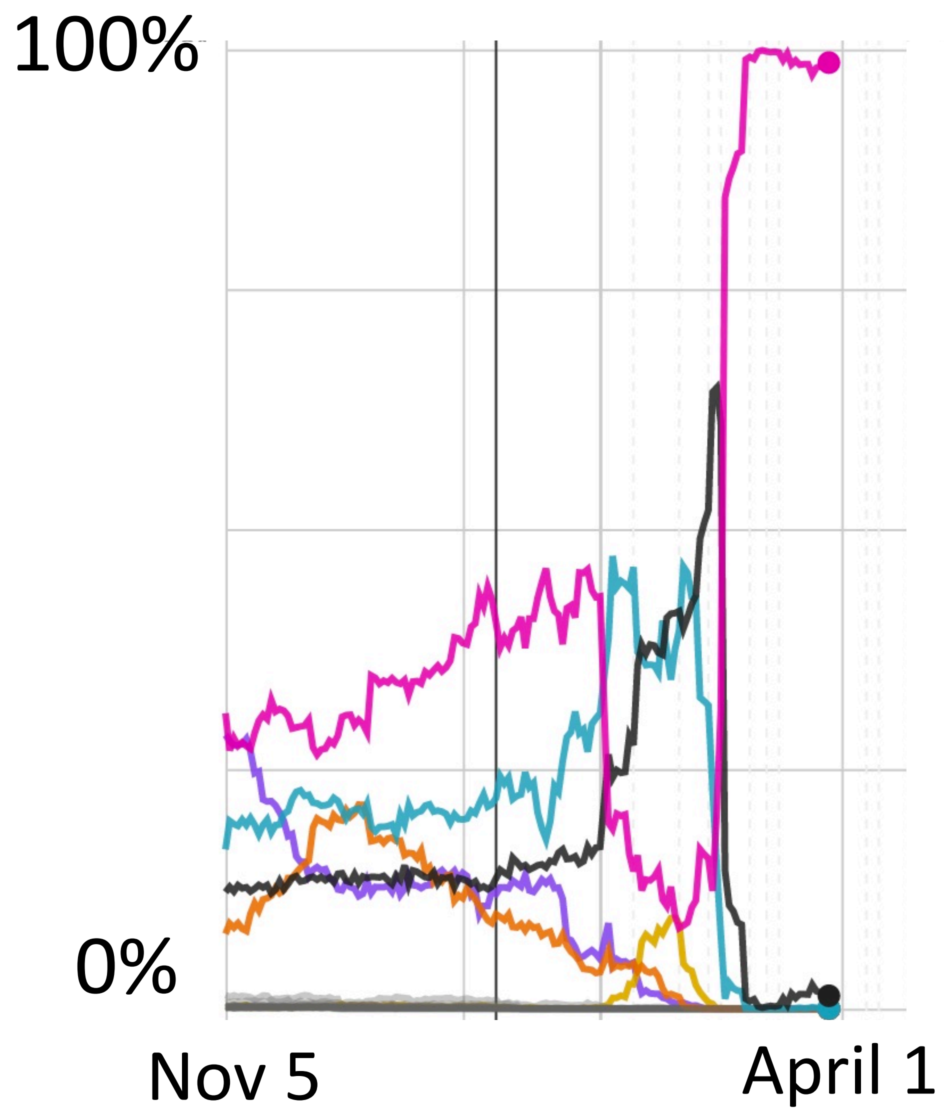
Bet against them.

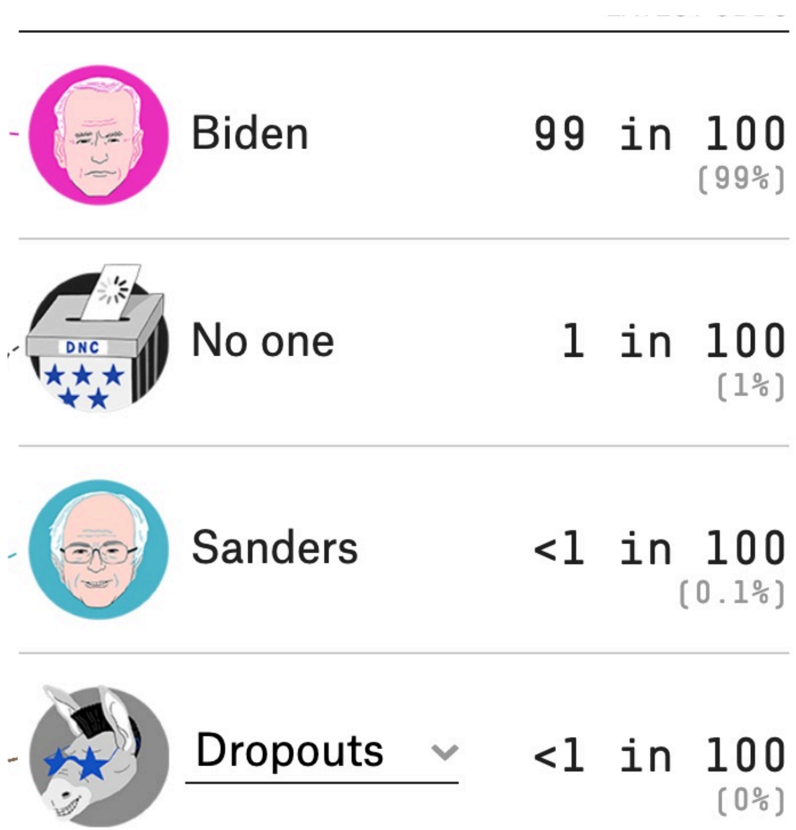
Test statistical hypotheses the same way.

- New use of likelihood ratios.
- Alternative to power.

Testing pundits and weather forecasters

Changing probabilities for Democratic candidate





Screen shot from fivethirtyeight.com on March 29

Predictions for the NBA (National Basketball Association) championship

	January 7	March 12
Bucks	25%	20%
Clippers	19%	26%
Lakers	17%	27%
76ers	17%	10%
Rockets	12%	7%
Raptors	3%	2%
Celtics	3%	6%
Nuggets	2%	1%
Mavericks	2%	<1%

March 12 was the last update before the season was suspended.

Testing by betting for statisticians

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.

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

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.

Successive tests of P

- P purports to describe $Y_1, Y_2, \ldots$.
- I test P by buying $S_1(Y_1)$ for \$1. Betting score $S_1(y_1)$ is mediocre — not much larger than 1.
- I continue testing. Score $S_2(Y_2)$ again mediocre.

Two ways of filling out the story

- I made the second bet by taking another \$1 out of my wallet. So I risked \$2. Final betting score is the mediocre

$$\frac{S_1(y_1) + S_2(y_2)}{2}.$$

- I made the second bet risking the winnings from the first. Final betting score is

$$S_1(y_1)S_2(y_2).$$

The second way is more powerful. So aim for large $S_1(y_1)S_2(y_2)$ rather than large $S_1(y_1) + S_2(y_2)$.

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

Three Examples

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 alternative.

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

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

$$y = 16.5$$

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

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

$y = 16.5$

- 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.

Warranties

- A $(1-\alpha)$-confidence set consists of all θ not rejected at level α.
- A $(1/\alpha)$-warranty set consists of all θ for which a strategy that always avoids bankruptcy does not multiply its initial capital by $1/\alpha$ or more.
- A warranty set can fail *a posteriori* in the same way a confidence set can.

A Glimpse at the game-theoretic foundations for probability

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

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

Ville's inequality. Suppose $Y_1, Y_2, \ldots$ is a stochastic proces, 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 $S_1, S_2, \ldots$ be the resulting capital process (nonnegative martingale). Then

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

Markov's inequality. If S is a nonnegative random variable, $E_P(S) = 1$, and $c > 0$, then

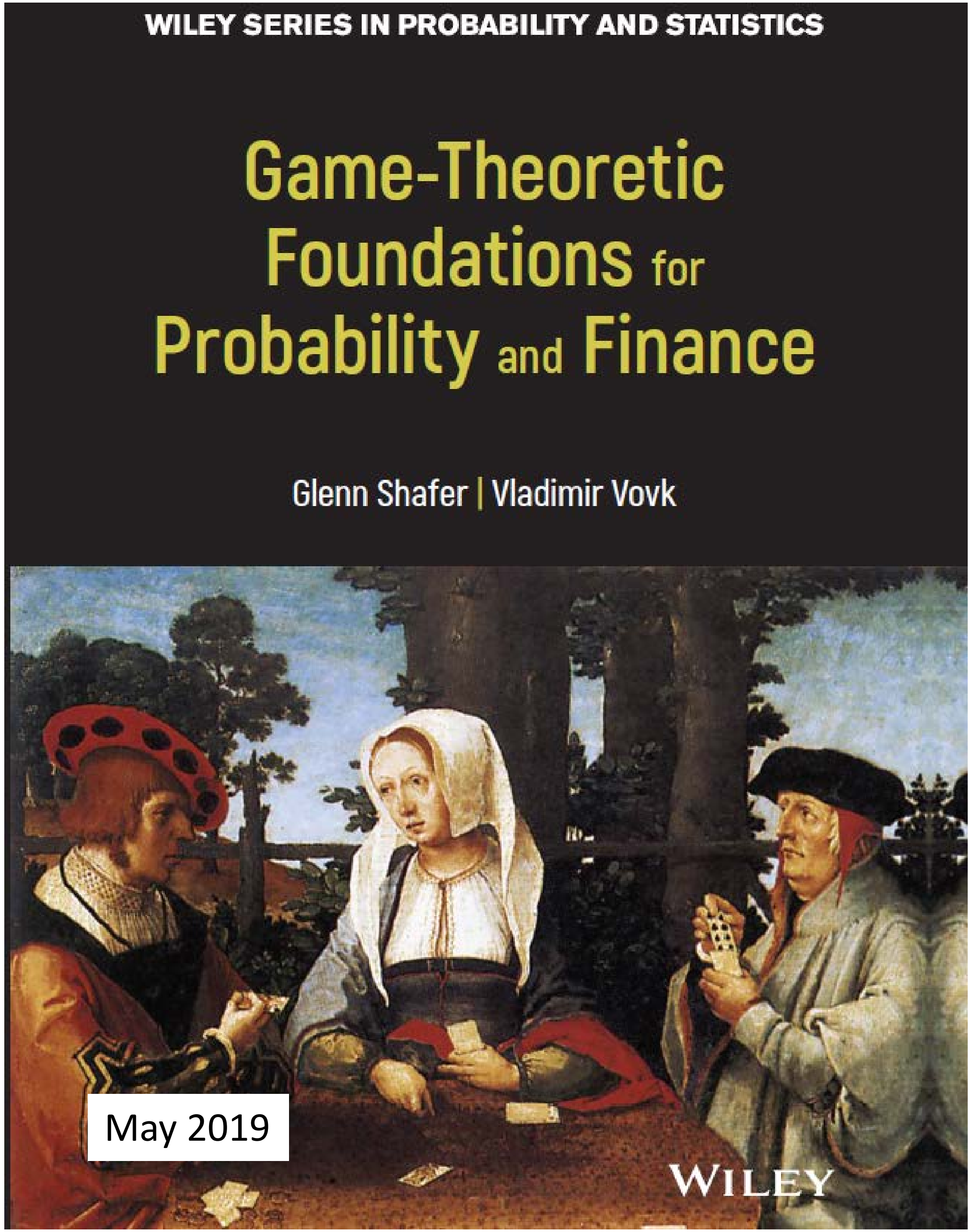
$$P(S \geq c) \leq \frac{1}{c}. \tag{1}$$

- "$P(E)$ is very small" is usually taken as a prediction that E will not happen. This gives empirical content to P.
- The inequality (1) is thus taken as predicting $S < c$.
- Another way of giving empirical content to P: A bet at P's odds will not multiply its capital by a large factor.
- Game-theoretic definition of probability: $\overline{P}(E) = p$ means that p is the least capital needed to 1 if E happens.

Vovk's inequality. Suppose you make successive bets starting with capital 1, not necessarily knowing what bets will be offered or having a strategy. Each time you bet so that your capital cannot become negative. Let $S_1, S_2, \ldots$ be the resulting capital process. Suppose $c > 0$. Then

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

where $\overline{P}(E) = p$ means that p is the least capital needed to play so that $\lim_{n\to\infty} S_n = 1$ if E happens.



Base mathematical probability on testing by betting.

Working papers at www.probabilityandfinance.com:

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