Let’s replace p-values with betting outcomes!

Glenn Shafer             Rutgers University
Brooklyn Quant Experience Lecture Series, NYU Tandon, 2 April 2020

How can we test the constantly fluctuating probabilities that Nate Silver offers for the outcomes of elections and sporting events? The natural (and perhaps only) way is to interpret the probabilities as betting offers and to bet against him. He fails our test if we multiply our money by a large factor.

We can test a statistical hypothesis, as well as the efficiency of a financial market, in the same way. In the case of statistical hypotheses, this leads to a new understanding of likelihood ratios and to an alternative to the notion of power.

Testing by Betting

Probability forecasts abound:
--weather, sports, elections, earnings

How can we test them?
This book bases the theory of probability itself on testing by betting.

Some recent working papers at www.probabilityandfinance.com go beyond the book.

See especially:
• 47 (on efficient markets)
• 54 (on statistical testing)
New world of probability forecasting

- Often forecasts are not based on statistical models with estimated parameters.
- They may come from physical models (as in hurricane forecasting) or neural networks.
- Each forecast may be on a different topic.

But we can still test by betting.
From Nate Silver’s [fivethirtyeight.com](https://fivethirtyeight.com)

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

<table>
<thead>
<tr>
<th>Team</th>
<th>January 7</th>
<th>March 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bucks</td>
<td>25%</td>
<td>20%</td>
</tr>
<tr>
<td>Clippers</td>
<td>19%</td>
<td>26%</td>
</tr>
<tr>
<td>Lakers</td>
<td>17%</td>
<td>27%</td>
</tr>
<tr>
<td>76ers</td>
<td>17%</td>
<td>10%</td>
</tr>
<tr>
<td>Rockets</td>
<td>12%</td>
<td>7%</td>
</tr>
<tr>
<td>Raptors</td>
<td>3%</td>
<td>2%</td>
</tr>
<tr>
<td>Celtics</td>
<td>3%</td>
<td>6%</td>
</tr>
<tr>
<td>Nuggets</td>
<td>2%</td>
<td>1%</td>
</tr>
<tr>
<td>Mavericks</td>
<td>2%</td>
<td>&lt;1%</td>
</tr>
</tbody>
</table>

Chance of winning NBA finals
How each candidate’s chances of winning more than half of pledged delegates have changed over time.
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?
Testing by betting

• Bob tests Alice buying random variables for the expected values she assigns them.

• If Bob begins with $1, risks no more than this, and walks away with $100, he puts a big dent in her reputation.

• Alice may plead bad luck, but she cannot claim success as a forecaster.

Bob is more frequentist than Bayesian.
Bayesians make bets on hypotheses—bets that are never settled.
Bob makes bets that are settled, and uses the outcomes like p-values.
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 does not need to risk real money.

He can bet with play money.

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

When he uses play money, he does not need a counterparty to his bets.

So 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.)
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.*

*Important detail: Bob is not allowed to borrow more money; otherwise he could “go martingaling”.
A theorem related to this understanding of testing by betting:

\[
\text{Markov’s inequality says that when } Z \text{ is a non-negative payoff with expected value } \mu,
\]
\[
P(Z \geq K\mu) \leq \frac{1}{K}.
\]

There is at best one chance in 100 that you can multiply your money by 100, 
*provided that you risk no more than this.*

The existing literature on testing market efficiency does not test in this way! The portfolios and strategies do not control the total risked.

Joint hypothesis problem
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* $\alpha$, 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 \( \alpha \), say 0.05.
- Choose \( E \) such that \( P(E) = 0.05 \).
- Reject \( P \) if \( y \in E \).

Betting interpretation: Bet on \( E \).
- Pay $1.
- If \( E \) does not happen, get back $0.
- If \( E \) happens, get back $20.
  - Then brag that you discredited \( P \).
  - You multiplied your money by a large factor.
  - What better evidence against \( P \) 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 $\alpha$ 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 $\alpha$ for which the test rejects.
- The smaller the p-value, the more evidence against $P$.

**Betting alternative:** Instead of an all-or-nothing bet (a bet that pays either $0$ or $20$, say), make a bet on $Y$ that can pay many different amounts.
- Such a bet is a function $S(Y)$.
- Choose $S$ such that $E_P(S) = 1$.
- Pay $1$.
- Get back $S(y)$. So $S(y)$ is the factor by which you multiplied your money.
- Call $S(y)$ your *betting score*.
- The larger $S(y)$, the more evidence against $P$. 

19
Choose $S$ such that $E_P(S) = 1$.

Pay $1$. Get back $S(y)$.

Your betting score $S(y)$ is the factor by which you multiply your money.

If $E_P(S) \neq 1$, the betting score is $\frac{S(y)}{E_P(S)}$.

The betting score does not change when we multiply $S$ by a positive constant.

You can bet so little that both $E_P(S)$ and $S(y)$ are negligible.

No decision theory here.

No need to play with real money.

It’s only a game!
Betting score
    = factor by which I multiply money risked.

Large betting score
    = best evidence I can have against $P$.

But maybe I was merely lucky.

Betting language
    = best way to communicate uncertainty.
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.
• My bet $S$ defines the alternative hypothesis $Q = SP$, even if I did not think about $Q$ when choosing $S$. (Perhaps I did not know the theory. Perhaps $Q$ is difficult to calculate.)

• On the other hand, if I begin with an alternative $Q$, then I can make the bet $Q/P$.

**Proof that $Q/P$ is a bet:** $E_P(Q/P) = 1$, because

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

But is liking $Q$ any reason to choose $Q/P$ as my bet?
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 $\mathbb{E}_Q(\ln S)$.

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

Kullback-Leibler divergence
Gibbs’s inequality

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

When $S$ is the product of successive factors, $\mathbb{E}(\ln S)$ 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 $\exp(E_Q(\ln S))$.

$$E_Q(\ln S) = \sum_y Q(y) \ln S(y) = \sum_y P(y) S(y) \ln S(y) = E_P(S \ln S)$$

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

<table>
<thead>
<tr>
<th>Proposed study</th>
<th>name</th>
<th>notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>initially unknown outcome</td>
<td>phenomenon</td>
<td>$Y$</td>
</tr>
<tr>
<td>probability distribution for $Y$</td>
<td>null hypothesis</td>
<td>$P$</td>
</tr>
<tr>
<td>nonnegative function of $Y$ with expected value 1 under $P$</td>
<td>bet</td>
<td>$S$</td>
</tr>
<tr>
<td>$S \times P$</td>
<td>implied alternative</td>
<td>$Q$</td>
</tr>
<tr>
<td>$\exp (E_Q(\ln S))$</td>
<td>implied target</td>
<td>$S^*$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>actual value of $Y$</td>
<td>outcome</td>
<td>$y$</td>
</tr>
<tr>
<td>factor by which money risked has been multiplied</td>
<td>betting score</td>
<td>$S(y)$</td>
</tr>
</tbody>
</table>
Three Examples
Many if not most misuses of p-values fall into one of three groups:

1. An estimate is statistically and practically significant but hopelessly contaminated with noise.

2. A test with a conventional significance level and high power against a very distinct alternative rejects the null hypothesis with a borderline outcome even though the null has greater likelihood than the alternative.

3. A high p-value is interpreted as evidence for the null hypothesis.
Here are elementary examples that show how the three cases can be handled with betting scores.

In each case, the test statistic is normal with standard deviation 10 under both the null and alternative distributions of the test statistic.
Example 1.

Estimate statistically and practically significant but hopelessly contaminated with noise.
Example 1. Suppose $P$ says that $Y$ is normal with mean 0 and standard deviation 10, $Q$ says that $Y$ is normal with mean 1 and standard deviation 10, and we observe $y = 30$.

- Statistician A simply calculates a p-value: $P(Y \geq 30) \approx 0.00135$. She concludes that $P$ is strongly discredited.

- Statistician B uses the Neyman-Pearson test with significance level $\alpha = 0.05$, which rejects $P$ when $y > 16.5$. Its power is only about 6%. Seeing $y = 30$, it does reject $P$. If she used the test as a bet, the statistician has multiplied the money she risked by 20.

- Statistician C uses the bet $S$ given by

$$S(y) := \frac{q(y)}{p(y)} = \frac{(10\sqrt{2\pi})^{-1} \exp(-(y-1)^2/200)}{(10\sqrt{2\pi})^{-1} \exp(-y^2/200)} = \exp \left( \frac{2y - 1}{200} \right),$$

for which

$$\mathbb{E}_Q(\ln(S)) = \mathbb{E}_Q \left( \frac{2y - 1}{200} \right) = \frac{1}{200},$$

so that the implied target is $\exp(1/200) \approx 1.005$. She does a little better than this very low target; she multiplies the money she risked by $\exp(59/200) \approx 1.34$.

The power and implied target both say the study is a waste of time.

The betting score of 1.34 confirms this.

The low p-value and Neyman-Pearson rejection give a misleading verdict in favor of $Q$. 
Example 2.

Test with conventional significance level and high power rejects the null even though the likelihood ratio favors the null.
Example 2. Now the case of high power and a borderline outcome: $P$ says that $Y$ is normal with mean 0 and standard deviation 10, $Q$ says that $Y$ is normal with mean 37 and standard deviation 10, and we observe $y = 16.5$.

- Statistician A again calculates a p-value: $P(Y \geq 16.5) \approx 0.0495$. She concludes that $P$ is discredited.

- Statistician B uses the Neyman-Pearson test that rejects when $y > 16.445$. This test has significance level $\alpha = 0.05$, and its power under $Q$ is almost 98%. It rejects; Statistician B multiplies the money she risked by 20.

- Statistician C uses the bet $S$ given by $S(y) := q(y)/p(y)$. Calculating as in the previous example, we see that $S$’s implied target is 939 and yet the betting score is only $S(16.5) = 0.477$. Rather than multiply her money, Statistician C has lost more than half of it. She concludes that the evidence from her bet very mildly favors $P$ relative to $Q$.

Assuming that $Q$ is indeed a plausible alternative, the high power and high implied target suggest that the study is meritorious. But the low p-value and the Neyman-Pearson rejection of $P$ are misleading. The betting score points in the other direction, albeit not enough to merit attention.
Example 3.

A high p-value is interpreted as evidence for the null.
Example 3. Now the case of a non-significant outcome: $P$ says that $Y$ is normal with mean 0 and standard deviation 10, $Q$ says that $Y$ is normal with mean 20 and standard deviation 10, and we observe $y = 5$.

- Statistician A calculates the p-value $P(Y \geq 5) \approx 0.3085$. As this is not very small, she concludes that the study provides no evidence about $P$.

- Statistician B uses the Neyman-Pearson test that rejects when $y > 16.445$. This test has significance level $\alpha = 0.05$, and its power under $Q$ is about 64%. It does not reject; Statistician B loses all the money she risked.

- Statistician C uses the bet $S$ given by $S(y) := q(y)/p(y)$. This time $S$’s implied target is approximately 7.39 and yet the actual betting score is only $S(5) \approx 0.368$. Statistician C again loses more than half her money. She again concludes that the evidence from her bet favors $P$ relative to $Q$ but not enough to merit attention.

In this case, the power and the implied target both suggested that the study was marginal. The Neyman-Pearson conclusion was “no evidence”. The bet $S$ provides the same conclusion; the score $S(y)$ favors $P$ relative to $Q$ but too weakly to merit attention.