Hypothesis Testing as a Game

JAAF Symposium 2018

IIM Ahmedabad

January 9, 2018

Glenn Shafer

1. History: [Fermat = measure theory] vs [Pascal = game theory]
2. Game-theoretic hypothesis testing
3. Calibrating p-values: $\sqrt{p}$
4. The game of p-hacking
5. Can empirical researchers document the game they play?
6. Can auditors document the game they play?

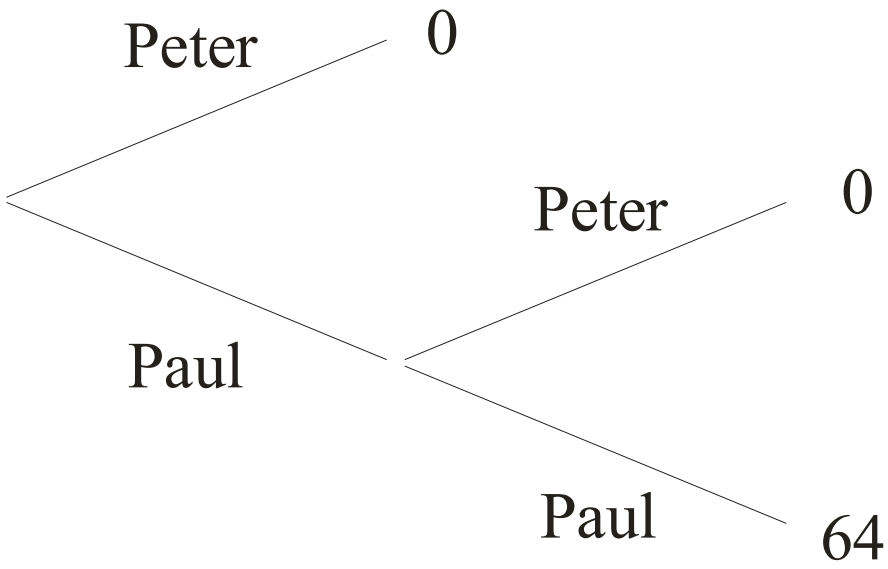
Fermat vs Pascal

In 1654, these two Frenchmen solved the division problem in different ways.

Pierre Fermat
1601-1665

Blaise Pascal
1623-1662

The division problem (aka problem of points)

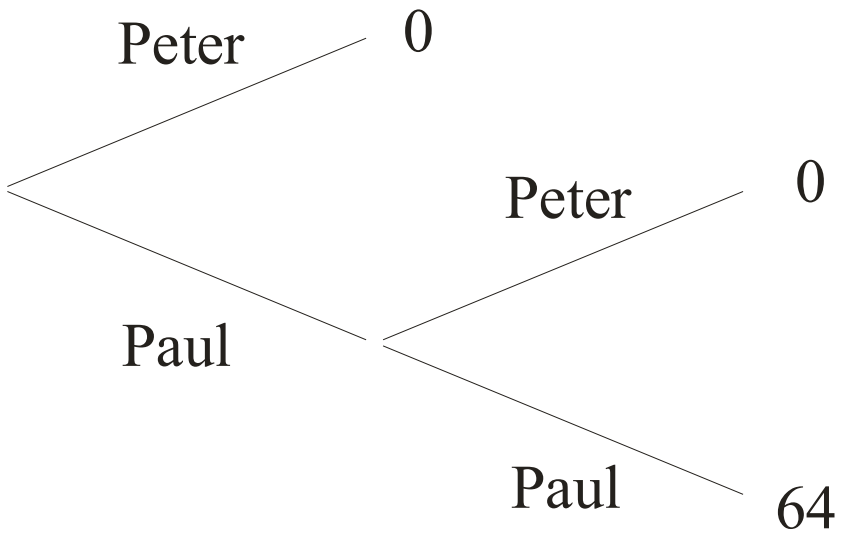


Paul's payoffs are shown.

Paul needs 2 points to win.
Peter needs only one.

If the game must be broken off, how should the 64 pistoles be divided?

Fermat's answer (measure theory)



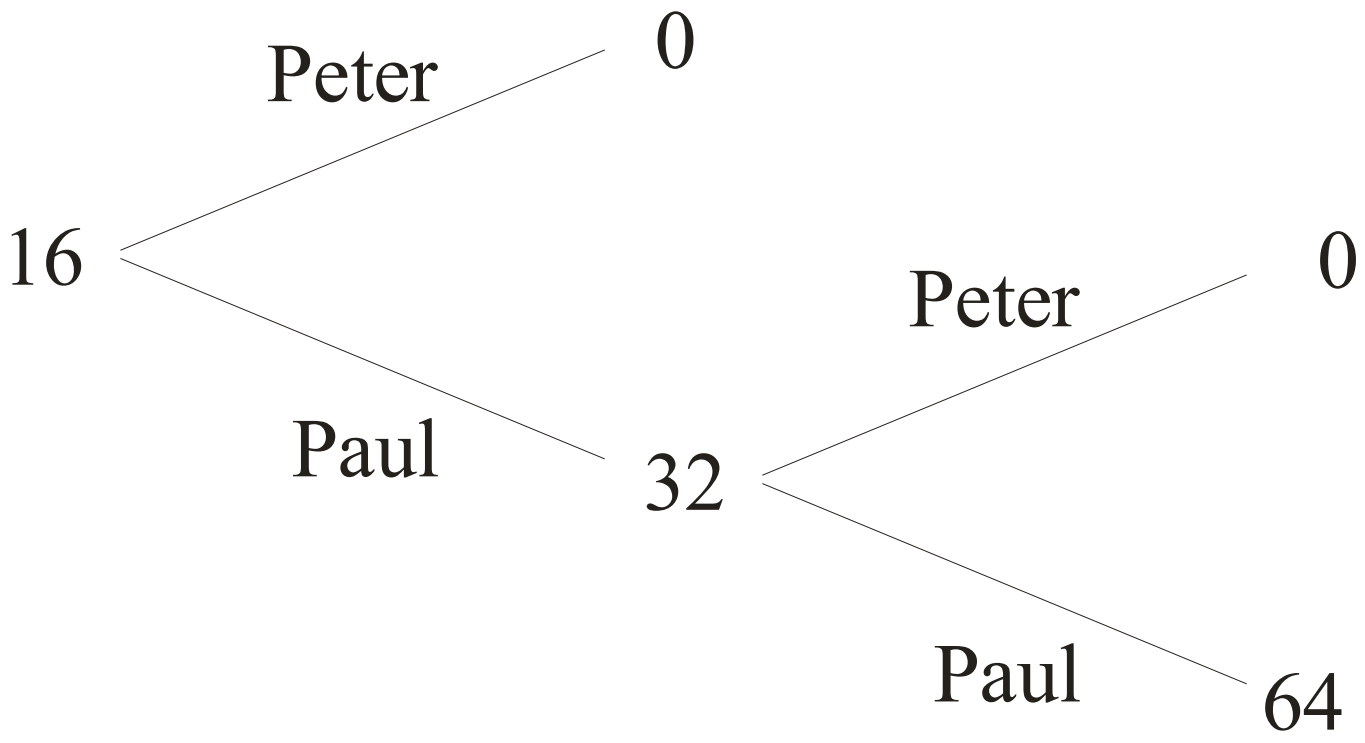
Suppose they play two rounds.

There are 4 possible outcomes:

1. Peter wins first, Peter wins second
2. Peter wins first, Paul wins second
3. Paul wins first, Peter wins second
4. Paul wins first, Paul wins second

Paul wins only in outcome 4. So his share should be ¼, or 16 pistoles.

Pascal's answer (game theory)



FERMAT

17th century France

PASCAL

De vetula, written in the 13th century and used in European universities into the 16th, explained the principles Fermat used.

The problem: Find the chances in dice games.

Pascal's solution was taught in some abacus schools in the 15th century. Others gave other solutions.

The problem: Settle unfulfilled business contracts or bets on interrupted ball games, archery competitions, chess tournaments, etc.

Pacioli's solution was in his book on business arithmetic, where he also introduced double-entry accounting.

FERMAT — 17th century France — PASCAL

13th century Paris
De vetula counted chances for dice.

16th century Italy
Pacioli settled contracts.

VIA CATHOLIC UNIVERSITIES

VIA ABACUS SCHOOLS

12th century Spain
Arabic to Latin

13th-15th century Italy
Arabic to Italian

Arab mathematics
Al-Khwārizmi

Greek geometry

Hindu reckoning & combinatorics

Akkadia

Trade/Writing/Dice

4,000 years ago

Game of *triga* depicted in Alfonso of Castile's *Book of Games* (1283).

Two players alternate throwing three dice.
To win, throw three of a kind, 15 or greater, or 6 or less.

Probability of winning on the first throw ≈ 19%.
Probability player who goes first wins ≈ 55%.

Measure-theoretic probability
vs.
game-theoretic probability

1. Definition of probability

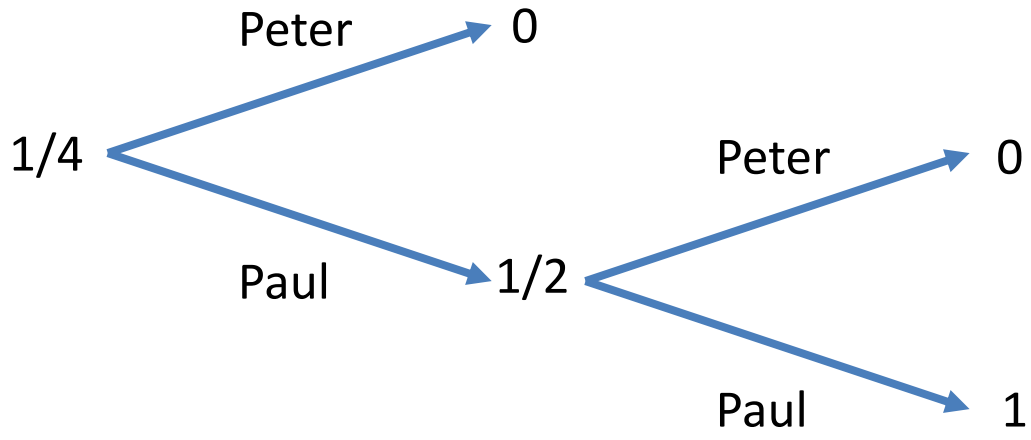
2. How to test a hypothesis

Measure-theoretic definition of probability

Peter wins first, Peter wins second
Peter wins first, Paul wins second
Paul wins first, Peter wins second
Paul wins first, Paul wins second

P(A) = fraction of equally likely cases favoring A

Game-theoretic definition of probability



P(A) = amount you must risk to get 1 if A happens

Measure-theoretic hypothesis testing

To test a probability model, choose an event *E* to which it assigns a small probability .

Reject at significance level α if *E* happens.

Game-theoretic hypothesis testing

To test a system of probabilities, bet at those probabilities.

Reject at significance level α if you multiply the money you risk by $1/\alpha$ or more.

Game-theoretic hypothesis testing

To test a system of probabilities, bet at those probabilities.

Reject at significance level α if you multiply the money you risk by $1/\alpha$ or more.

Game-theoretic testing of market efficiency

To test a system of prices, trade at those prices.

Reject at significance level α if you multiply the money you risk by $1/\alpha$ or more.

Neyman-Pearson is game-theoretic, but p-values are not.

A Neyman-Pearson significance level α IS a legitimate game-theoretic significance level.

When you choose the event *E* that has probability α according to the theory, bet \$1 that *E* will happen. You will turn the \$1 you risk into \$1/ α if *E* happens.

The p-value from a test statistic *T* IS NOT a legitimate game-theoretic significance level.

The p-value *p* is the probability the model gives to $T \geq t$, where *t* is the observed value of *T*. It has the property

$$P(p \leq \alpha) \leq \alpha.$$

We did not choose the event $E = \{p \leq \alpha\}$ in advance and so did not bet on it. Pretending we did is cheating a little.

(Only a little because we did choose *T* in advance.)

Calibrating a p-value

Why do we like p-values? Why not set a 5% level in advance and stick to it?

Because we want to recognize that there is even stronger evidence if the test comes out even more significant.

To accommodate this desire game-theoretically, specify in advance several levels of significance and distribute our betting money among them.

Example: Bet \$1 each on significance levels 5%, 1%, and 0.1%.

- If $p>0.05$, we turned \$3 into \$0. No evidence.
- If $0.01<p\leq 0.05$, we turned \$3 into \$20. Significance level = 3/20=0.15.
- If $0.001<p\leq 0.01$, we turned \$3 into \$120. Significance level = 3/120=0.025.
- If $p\leq 0.001$, we turned \$3 into \$1120. Significance level = $3/1120 \approx 0.0027$.

More sophisticated: distribute your capital continuously over all the significance levels between 0 and 1.

Simple rule of thumb: $\sqrt{p}$

Spread one unit of capital over $(0,1)$ using the density $\frac{1}{2\sqrt{x}}$. Bet each increment $\frac{1}{2\sqrt{x}}dx$ on $\{p \le x\}$, turning it into

$$\frac{1}{x}\,\frac{1}{2\sqrt{x}}dx$$

if $\{p \le x\}$ happens. The overall factor by which you multiply your money is

$$\int_p^1 \frac{1}{x}\frac{1}{2\sqrt{x}}dx = \frac{1}{\sqrt{p}} - 1$$

Significance level = inverse of factor =

$$\frac{\sqrt{p}}{1-\sqrt{p}} \approx \sqrt{p}.$$

If we treat p-values larger than 0.25 as "no evidence", spreading our initial capital over $(0, 0.25)$ only, we decrease the significance level to $\sqrt{p}/2$.

Proposals on the table

- 2 standard deviations $\rightarrow$ 3 standard deviations

  Campbell R. Harvey, "Presidential Address: The scientific outlook in financial economics", *Journal of Finance* 72(4):1399-1440, August 2017, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2893930

- 0.05 $\rightarrow$ 0.005.

  Manifesto by 72 prominent mathematical statisticians, mostly Bayesians, July 2017, https://psyarxiv.com/mky9j/

			$\sqrt{p}$ rule		$\sqrt{p}/2$ rule	
z	$\frac{1}{p}$	p	$\sqrt{\frac{1}{p}-1}$	$\frac{1}{\sqrt{\frac{1}{p}-1}}$	$\frac{2}{\sqrt{p}}$	$\frac{\sqrt{p}}{2}$
2.0	20	0.05	3.5	0.29	8.9	0.11
2.8	200	0.005	13	0.076	28	0.035
3.0	370	0.0027	18	0.055	38	0.026
3.9	10,000	0.0001	99	0.010	200	0.005

Changing the criteria for publication from 2σ to 3σ or from 5% to 0.5% makes the game harder (and the journal editor's life easier) but does not change the nature of the game.

Everyone knows that the real problem is multiple testing, or p-hacking.

Shifting from 2σ to 3σ or from 5% to 0.5% roughly corrects for not fixing the significance level in advance, leaving the problem of p-hacking untouched.

Because p-hacking is a game, we can evaluate its results only when we see the play of the game.

Radical proposal: Require authors to document p-hacking and base an argument on this documentation.

Antoine Augustin Cournot explained the p-hacking game in 1843 in §111 of his *Exposition de la théorie des chances et des probabilités.*

. . . Suppose, for example, that we want to determine, on the basis of a large number of observations collected in a country like France, the chance of a masculine birth. We know that in general it exceeds 1/2. We can first distinguish between legitimate births and those outside marriage, and as we will find, with large numbers of observations, a very appreciable difference between the values of the ratio of masculine births to total births, depending on whether the births are legitimate or illegitimate, we will conclude with very high probability that the chance of a masculine birth in the category of legitimate births is appreciably higher than the chance of the event in the category of births outside marriage. We can further distinguish between births in the countryside and births in the city, and we will arrive at a similar conclusion. These two classifications come to mind so naturally that they have been an object for examination for all statisticians.

. . . we could also classify births according to their order in the family, according to the age, profession, wealth, and religion of the parents. . . [A]s the number of groupings thus grows without limit, it is more and more likely *a priori* that merely as a result of chance at least one of the groupings will produce values appreciably different in the two distinct categories.

Consequently, . . . for a statistician who undertakes a thorough investigation, the probability of a deviation of given size not being attributable to chance will have very different values depending on whether he has tried more or fewer groupings.

. . . But usually the groupings that the experimenter went through leave no trace; the public only sees the result that seemed to merit being brought to its attention. Consequently, an individual unacquainted with the system of groupings that preceded the result will have absolutely no fixed rule for betting on whether the result can be attributed to chance. There is no way to give an approximate value to the ratio of erroneous to total judgments a rule would produce. . .

- Oscar Sheynin's translation of Cournot's book is available at http://sheynin.de/download/cournot.pdf.
- See also my working paper "Cournot in English" at http://www.probabilityandfinance.com/articles/48.pdf.

Is it possible for empirical researchers to document their search process in published articles?

Can we tell the truth about

- How many ways we massaged the data?
- How many hypotheses we tried?

Can our colleagues believe us?

Would requiring the effort make other modes of research more competitive?

Can the record of an audit engagement be used to make a game-theoretic argument?

Here multiple testing can be a virtue.

But the cogency of the audit may benefit from explicit models for the hypotheses of material error the auditor is testing.

References

1. *Probability and Finance: It's only a game*, by Glenn Shafer and Vladimir Vovk. Wiley, 2001.

   See www.probabilityandfinance.com.

2. Game-theoretic significance testing, by Glenn Shafer. http://www.probabilityandfinance.com/articles/49.pdf.

   Game-theoretic probability gives us a new way to think about the problem of adjusting p-values to account for multiple testing and provides concrete rules for adjusting and combining p-values.

3. Marie-France Bru and Bernard Bru on dice games and contracts, by Glenn Shafer. To appear in *Statistical Science*.

   Counting chances for dice and estimating fair price came together in Fermat and Pascal's 1654 correspondence on dividing the stakes in a prematurely halted game. Fermat used centuries-old principles for analyzing dice games, while Pascal used centuries-old principles of commercial arithmetic.