How game-theoretic probability can reform theoretical statistics

Safe, Anytime-Valid Inference (SAVI) and Game-theoretic Statistics

Eurandom,

May 31, 2022

Glenn Shafer, Rutgers University

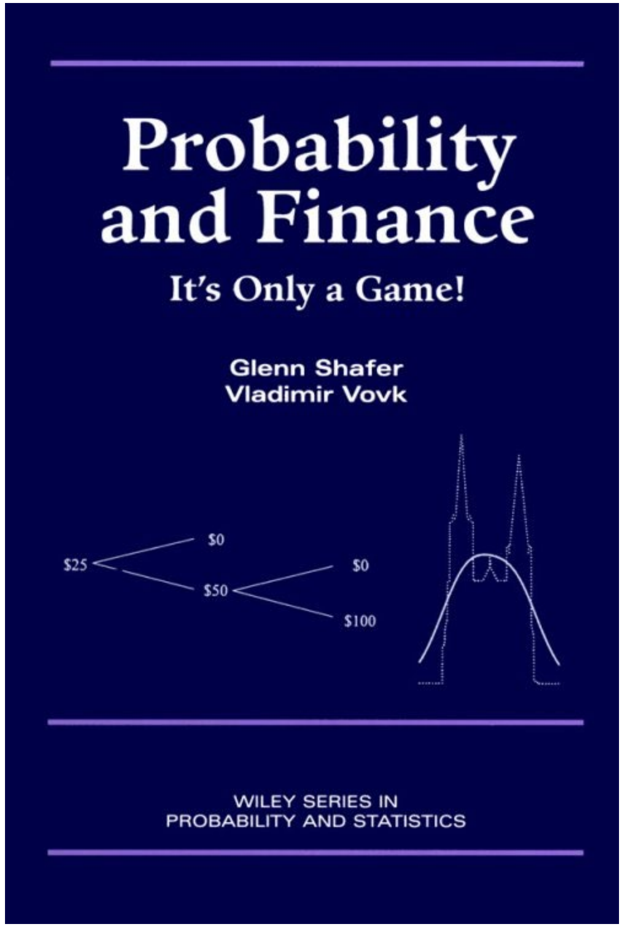
The Splendors and Miseries of Martingales

Their History from the Casino to Mathematics

Edited by Laurent Mazliak and Glenn Shafer, Birkhäuser, 2022

The nineteen chapters include:

- ROGER MANSUY, The Origins and Multiple Meanings of *Martingale*
- GLENN SHAFER, Martingales at the Casino
- BIENVENU/SHAFER/SHEN, Martingales in the Study of Randomness
- T. S. LAI, Encounters with Martingales in Statistics and Stochastic Optimization
- AALEN/ANDERSEN/BORGAN/GILL/KEIDING, Martingales in Survival Analysis
- TYRONE DUNCAN, Encounters with Martingales in Stochastic Control


Probability and Finance

It's Only a Game!

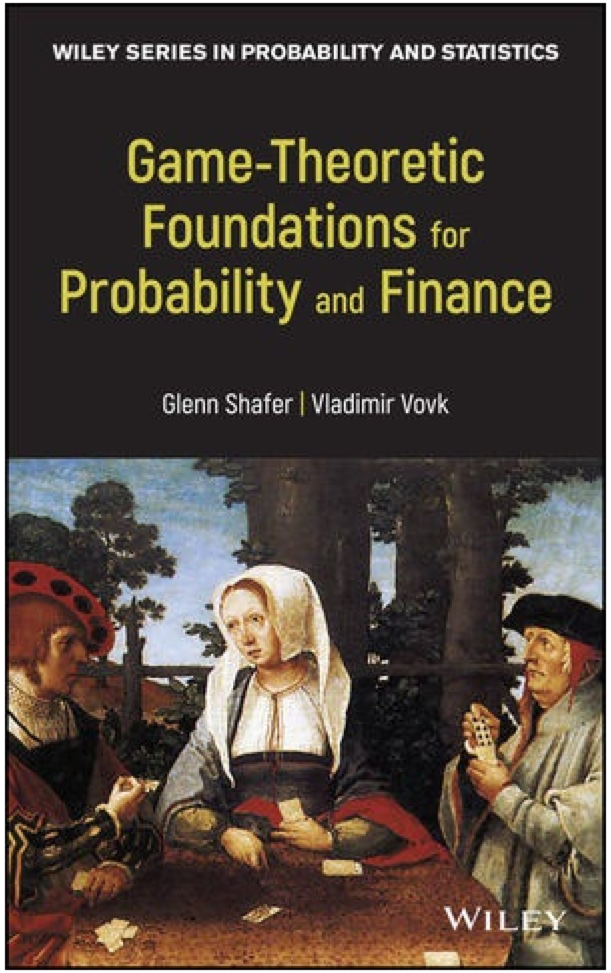
Glenn Shafer
Vladimir Vovk

WILEY SERIES IN
PROBABILITY AND STATISTICS

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.


WILEY SERIES IN PROBABILITY AND STATISTICS

Game-Theoretic Foundations for Probability and Finance

Glenn Shafer | Vladimir Vovk

WILEY

2019

- Puts game-theoretic probability on a par with discrete-time measure-theoretic probability as abstract theory.
- Applications (parametric and nonparametric confidence sequences, defensive forecasting, decision-making, CAPM, equity premium, stochastic calculus, calibration, etc.)

Game-theoretic probability

For *n* = 1,2,…

Forecaster announces *n*th probability.
Skeptic makes *n*th bet.
Reality announces *n*th outcome.

Probability defined by what bets can accomplish.

P(A) = 1/(max betting score on assumption A does not happen)

All three players are free agents. **Strategy defined in advance not required.**

Optional continuation is intrinsic in this setup. No need for special pleading.

Game-theoretic statistics

(Shafer/Vovk 2019, Chapter 10, **Using Testing Protocols in Science and Technology**)

For $n = 1,2,\ldots$

Reality announces parameter θ_n.
Forecaster announces p_n.
Skeptic makes bet $\mathrm{B}_n(y)$.
Reality announces outcome y_n.

- Players see each other's moves.
- Statistician is not a player and only sees *y*s.

Statistician tells Forecaster and Skeptic what to do as function of previous moves, including those Statistician does not see.

Can "e-processes" be defined in this setup?

How game-theoretic probability can reform theoretical statistics

1. **Prediction**

2. **Description**

3. **Protocol**

As presently taught, theoretical statistics falls short of 21st-century needs.

1. Prediction. As Leo Breiman taught us, our culture of modeling distances us from modern prediction.

2. Description. As David Freedman taught us, we produce investigators who treat convenience samples and entire populations as random samples.

3. Protocol. We calculate p-values and “condition” probabilities without really having a sampling plan.

1. Prediction

As Leo Breiman taught us, the culture of modeling distances theoretical statistics from modern prediction.

- It teaches practitioners to test models, not forecasters.
- It emphasizes methods of estimation irrelevant to the physical models used in weather prediction and the neural networks prized in machine learning.
- It emphasizes **iid** or exchangeable observations, which hardly exist in nature.

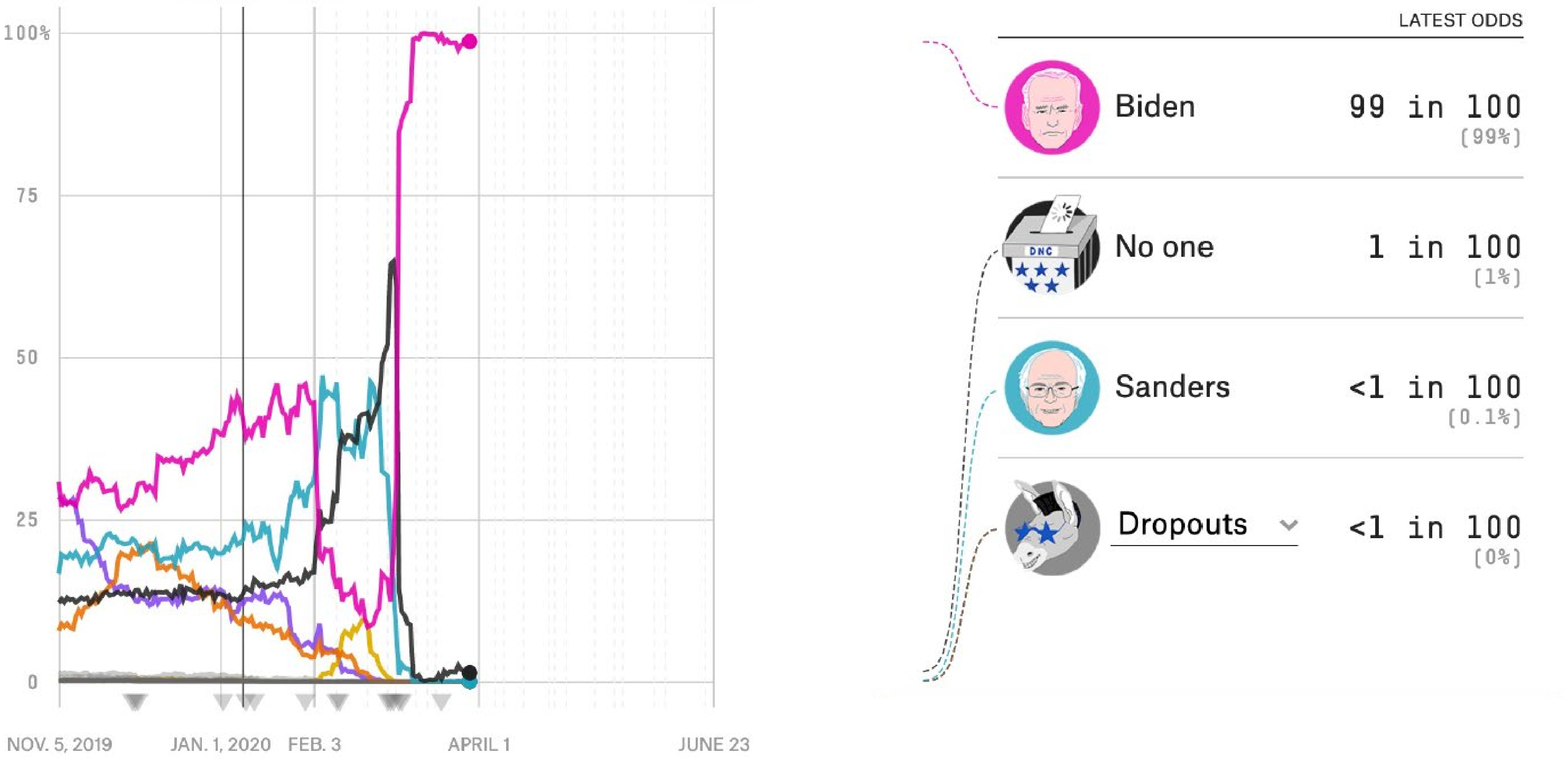
The first class in a course on theoretical statistics should be about testing probability forecasters.

Test probabilities for rain.

Test probabilities for elections.

Test financial analysts' predictions.

The statistics we teach millions of students each year gives no way or understanding or testing the changing probability predictions these students consume.



	LATEST ODDS
Biden	99 in 100 (99%)
No one	1 in 100 (1%)
Sanders	<1 in 100 (0.1%)
Dropouts	<1 in 100 (0%)

When we teach estimation in parametric models, we overemphasize **iid**, because stochastic processes are so complicated.

When we teach prediction, the special case of constant prediction seems uninteresting.

Always the Achilles' heel…

iid was central to the discredit and decline of Laplacean methods during the 19th century.

Laplace himself acknowledged that observational data is never **iid**:

> The law of exponentials that Mr. Gauss adopted as the law of the errors of observations probably hardly ever occurs in nature, which surely follows no constant law, but in which the laws of errors vary with the nature of the observational instruments and all the circumstances that accompany them.

--Letter written in 1816 to Bernhard August von Lindenau.

(Letter 899 in Roger Hahn’s *Correspondance de Pierre Simon Laplace (1749—1827)*, Brepols, Turnhout, Belgium, 2013)

But **iid** gained new life in Karl Pearson’s biometrics.

Randomly sample

- flowers from a field,
- crabs from the seashore.

Pearson’s sampling still lurks in our vision of theoretical statistics.

Even though the English school quickly expanded to observational data – e.g. Yule’s study of pauperism.

Look how Fisher took **iid** for granted:

> … there is no falsehood in interpreting any set of independent measurements as a random sample from an infinite population; for any such set of numbers are a random sample from the totality of numbers produced by the same matrix of causal conditions...

“On the Mathematical Foundations of Theoretical Statistics”, 1922, p. 313.

Fisher continued:

> As regards problems of specification, these are entirely a matter for the practical statistician … [W]e may know by experience what forms are likely to be suitable, and the adequacy of our choice may be tested *a posteriori*.

Really?

As Breiman and Freedman taught us, goodness-of-fit testing is useless for multiple regression and other complicated models.

2. Description

As David Freedman taught us, we produce investigators in the social sciences who treat convenience samples and entire populations as random samples.

- The best authors deplore this practice but provide no satisfying alternatives.

- We provide no way to understand the precision of parameters when a fitted model is purely descriptive.

The original sin of theoretical statistics

Treating convenience samples and entire populations as random samples

- In 1826, Joseph Fourier used a convenience sample to calculate confidence intervals for “the length of a masculine generation in France”.

- In the 1830s, Siméon Denis Poisson used census data to test hypotheses about the birth ratio in France.

Joseph Fourier

- Revolutionary. Administrator for Napoleon.
- After Napoleon fell, former student appointed him to a position at the census.
- He used 18th century records to estimate the length of "the masculine generation in France".

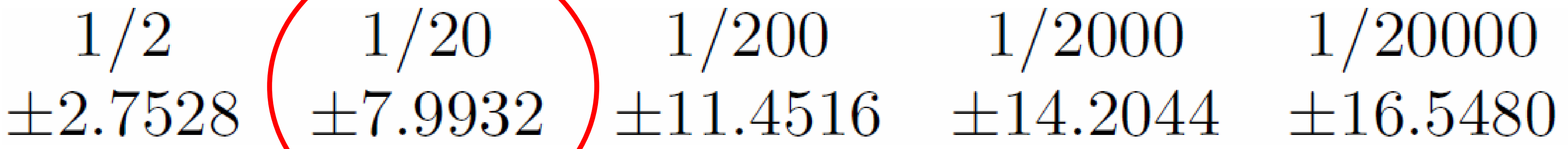
1768-1830

Masculine generation = average time, for fathers of sons, from father's birth to birth of first son.

Masculine generation = average time, for fathers of sons, from father's birth to birth of first son.

On the basis of 505 cases, Fourier estimated it to be 33.31 years.

Bounds on the error for five different probabilities:



1/2	1/20	1/200	1/2000	1/20000
±2.7528	±7.9932	±11.4516	±14.2044	±16.5480

months

95% confidence

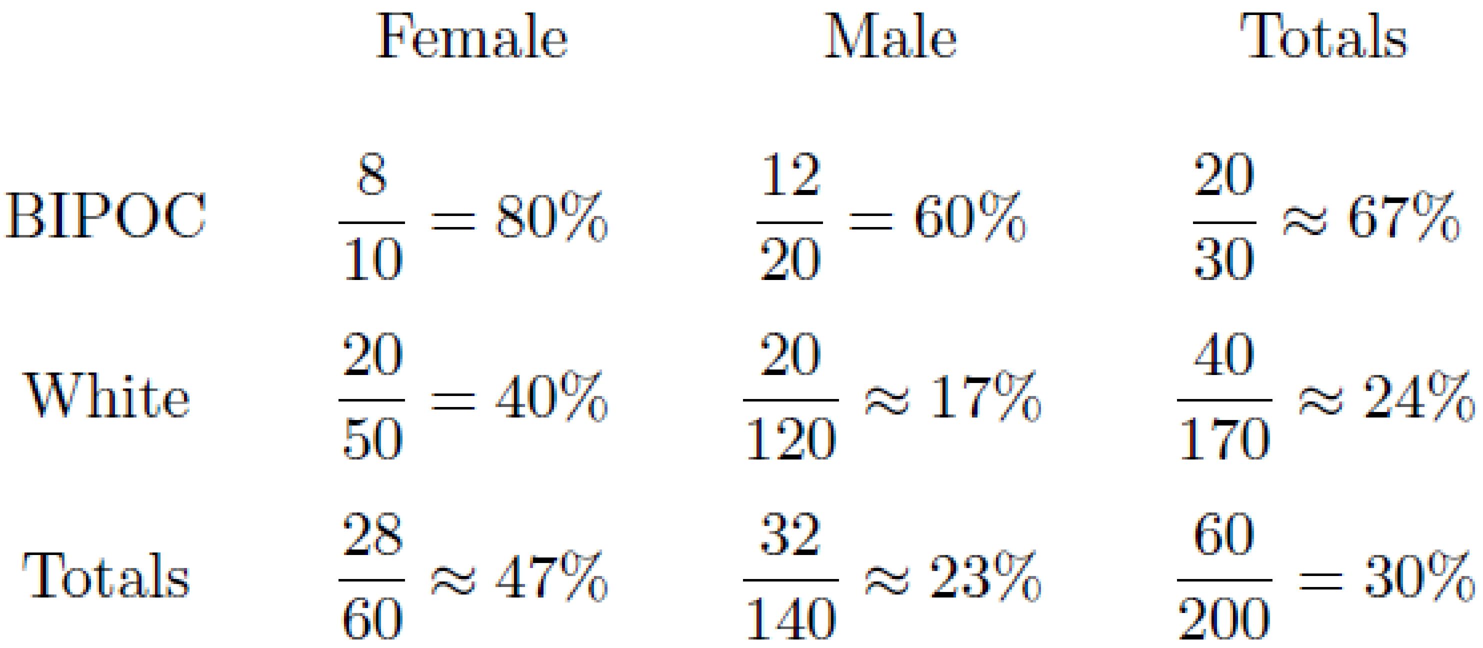
Convenience sample, from Paris records.

Very, very often, our “stochastic models” are phony, and the phoniness is unwanted.

Length of the masculine generation?

This is just a question of description.

We ask the 200 employees of our organization, “Have you experienced discrimination because of your identity?”



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

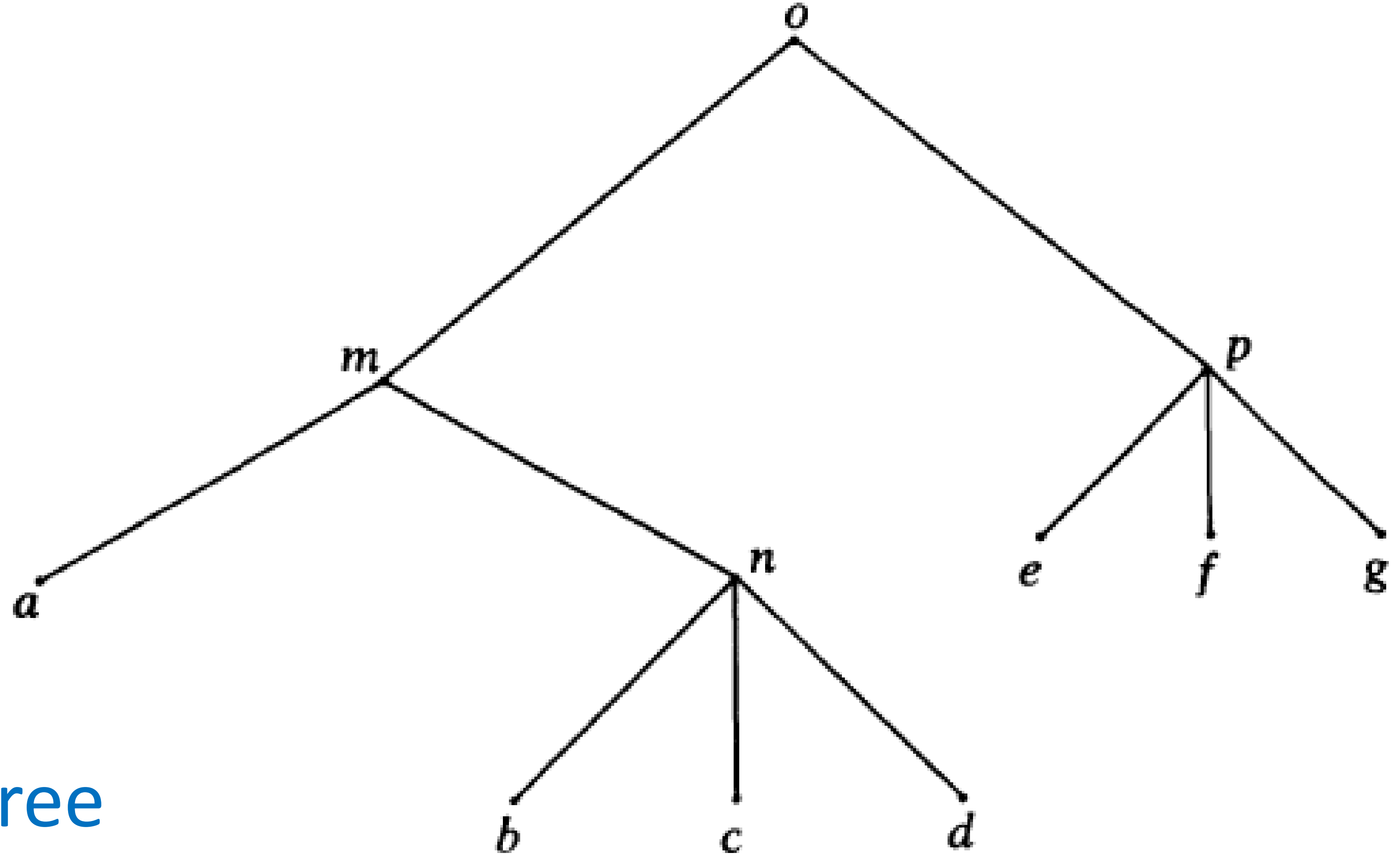
Working paper 59 at www.probabilityandfinance.com

3. **Protocol**

While paying lip service to measure-theoretic probability, we pretend that “conditioning” is legitimate without a filtration, sampling plan, or event tree.

This is the <u>modern</u> Bayesian sin.

modern = after 1950



Tree

$abcdefg$

$abcd|efg$

$a|bcd|e|f|g$

$a|b|c|d|e|f|g$

Filtration

Doob added filtrations to Kolmogorov’s axioms.

But still an advanced topic for statistical teaching.

Earliest occurrence of “conditionalize” in Google Scholar:

Estes and Suppes, 1957

http://suppes-corpus.stanford.edu/techreports/IMSSS_16.pdf

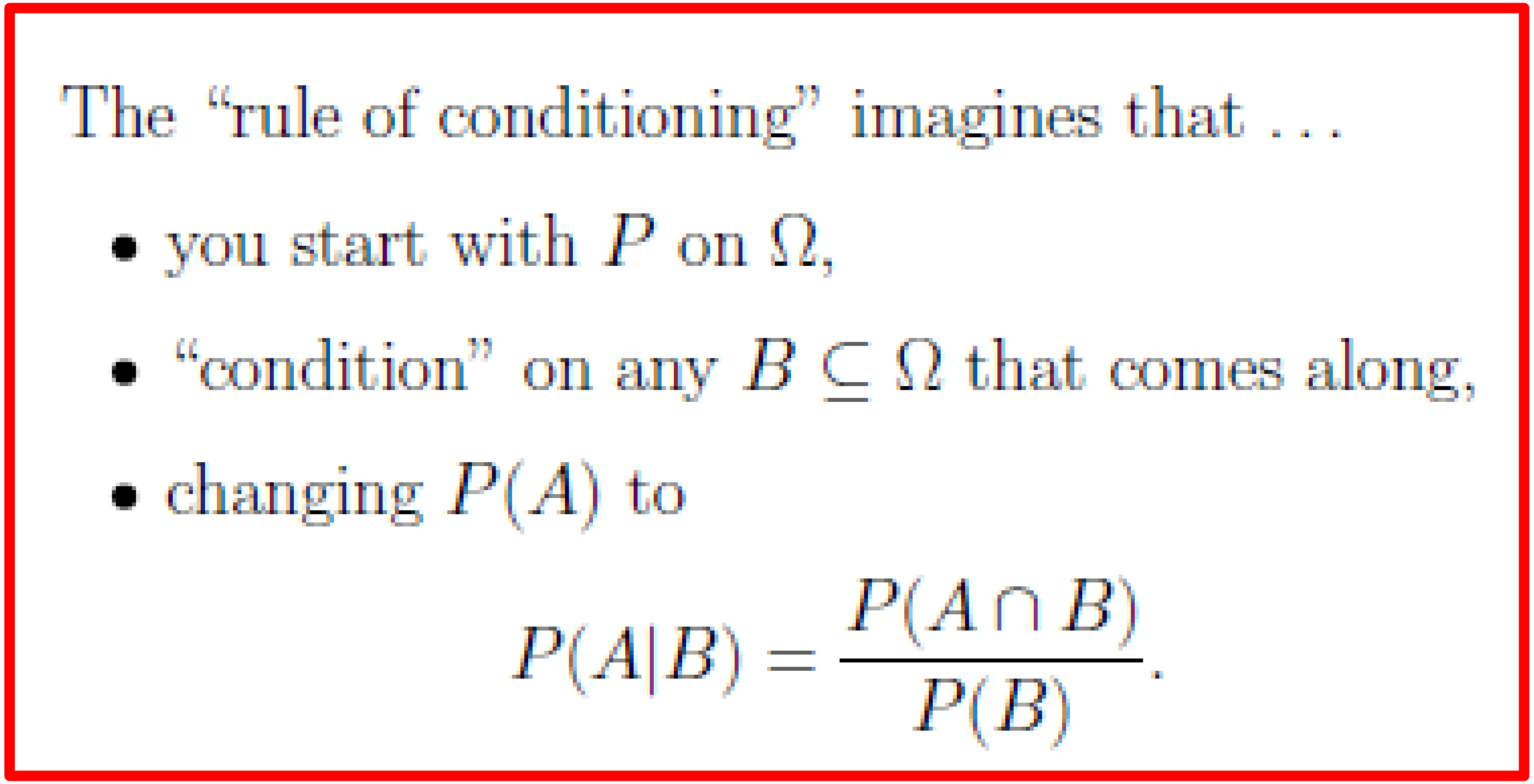
… the experimenter may conditionalize the probabilities of reinforcement upon preceding events of the sample space in whatever manner he pleases.

Earliest occurrences of “rule of conditioning” in Google Scholar and Google Books:

Glenn Shafer’s 1973 doctoral dissertation.

Glenn Shafer’s 1976 book, *A Mathematical Theory of Evidence*

Glenn Shafer’s 1976 article, “A theory of statistical evidence”



The “rule of conditioning” imagines that . . .

- you start with P on Ω,
- “condition” on any $B \subseteq \Omega$ that comes along,
- changing $P(A)$ to

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

This is …

- **foreign to the thinking of Bayes and Laplace**
- **and WRONG!**

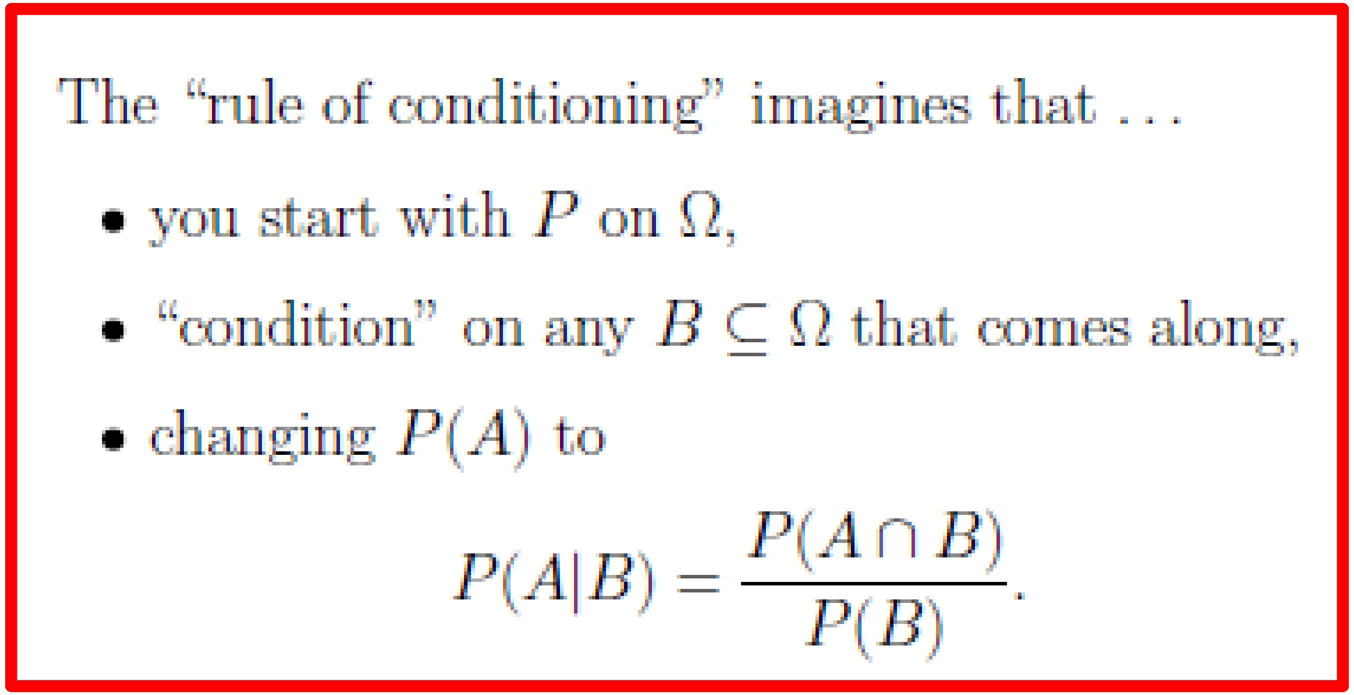
The "rule of conditioning" imagines that …

- you start with P on Ω,
- "condition" on any $B \subseteq \Omega$ that comes along,
- changing $P(A)$ to

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

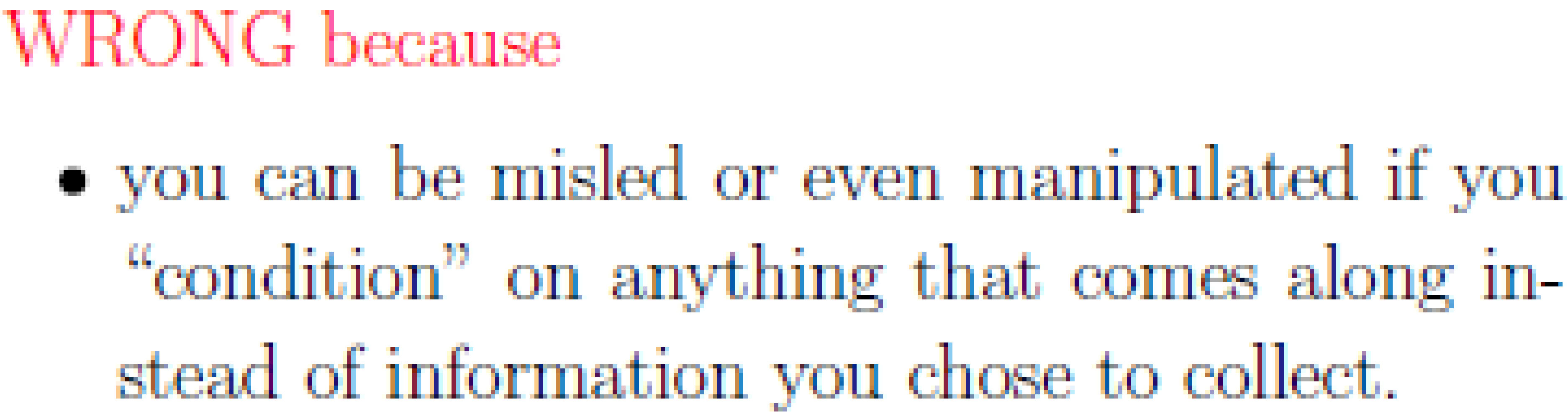
Foreign to classical Bayesian statisticians …

- …classical Bayesian statisticians were statisticians and so had a sampling plan,
- …the statistician has probabilities for (θ, x) and plans to observe x.



The "rule of conditioning" imagines that …

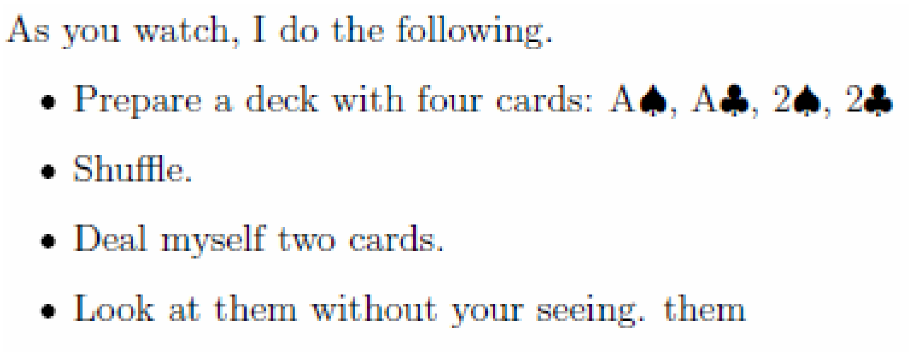
- you start with P on Ω,
- "condition" on any $B \subseteq \Omega$ that comes along,
- changing $P(A)$ to
$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$



WRONG because

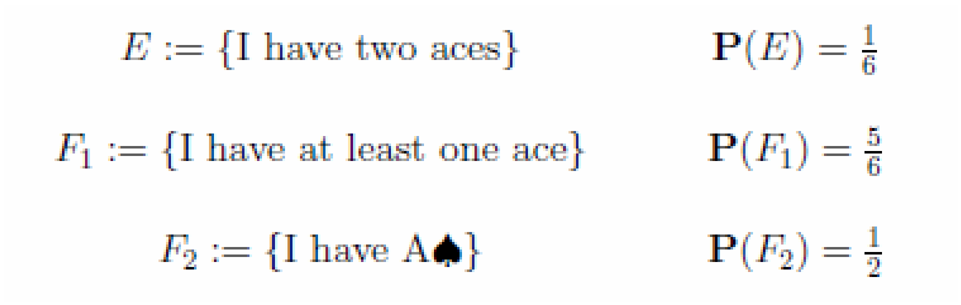
- you can be misled or even manipulated if you "condition" on anything that comes along instead of information you chose to collect.

Whitehead’s 1938 puzzle of the two aces



As you watch, I do the following.

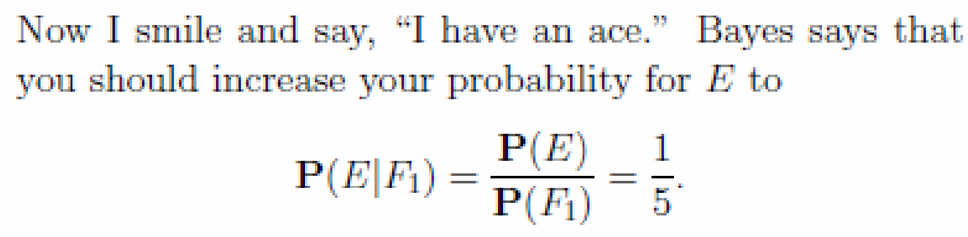
- Prepare a deck with four cards: A♠, A♣, 2♠, 2♣
- Shuffle.
- Deal myself two cards.
- Look at them without your seeing. them



$$E := \{\text{I have two aces}\} \qquad \mathbf{P}(E) = \tfrac{1}{6}$$

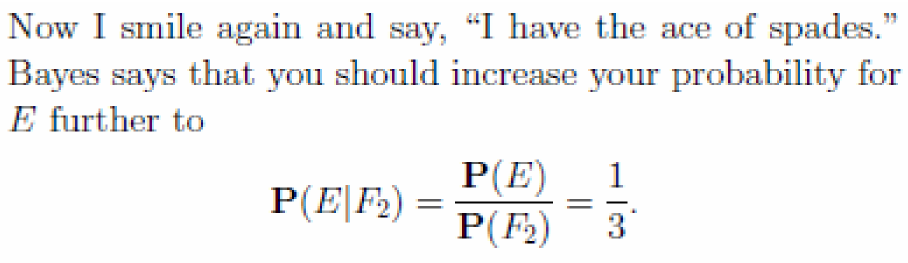
$$F_1 := \{\text{I have at least one ace}\} \qquad \mathbf{P}(F_1) = \tfrac{5}{6}$$

$$F_2 := \{\text{I have A}\spadesuit\} \qquad \mathbf{P}(F_2) = \tfrac{1}{2}$$



Now I smile and say, “I have an ace.” Bayes says that you should increase your probability for E to

$$\mathbf{P}(E|F_1) = \frac{\mathbf{P}(E)}{\mathbf{P}(F_1)} = \frac{1}{5}.$$



Now I smile again and say, “I have the ace of spades.” Bayes says that you should increase your probability for E further to

$$\mathbf{P}(E|F_2) = \frac{\mathbf{P}(E)}{\mathbf{P}(F_2)} = \frac{1}{3}.$$

Why should my identifying a suit change your probability?

You do not know why I gave you the information.

You need a larger model that includes probabilities for my behavior.

John E. Freund, Puzzle or paradox, *American Statistician* 19(4):29-44, 1965.
N. T. Gridgeman, Letter to the editor, *American Statistician* 21(3):38, 1967.
Glenn Shafer, A subjective interpretation of conditional probability. *Journal of Philosophical Logic* **12** 453-466. 1983

Neo-Bayesian philosophers “solve” the problem with …

The principle of total evidence

Condition on everything you have learned.

When you learn B, you also learn

C = {you learned B in a particular way}.

Condition on C.

Problem:

- C not in Ω unless Ω includes a sampling plan.
- In real life, you do not have such a plan.

1. **Prediction**

2. **Description**

3. **Protocol**

1. **Prediction**

2. **Description**

3. **Protocol**

Game-theoretic probability addresses the three issues in a natural way.

Game-theoretic probability addresses the three issues in a natural way.

1. Prediction. Test probability forecasters without regard to whether they are statistical models, physical models, neural nets, or human beings.

2. Description. Use parametric models, including regression models, descriptively by having different parameter values bet against each other.

3. Protocol. Betting makes a plan unavoidable. You cannot bet on what you do not plan to observe.

1. Prediction

Test probability forecasters without regard to whether they are statistical models, physical models, neural nets, or human beings.

By betting, you can test probability forecasts that change daily or even hourly (weather, elections).

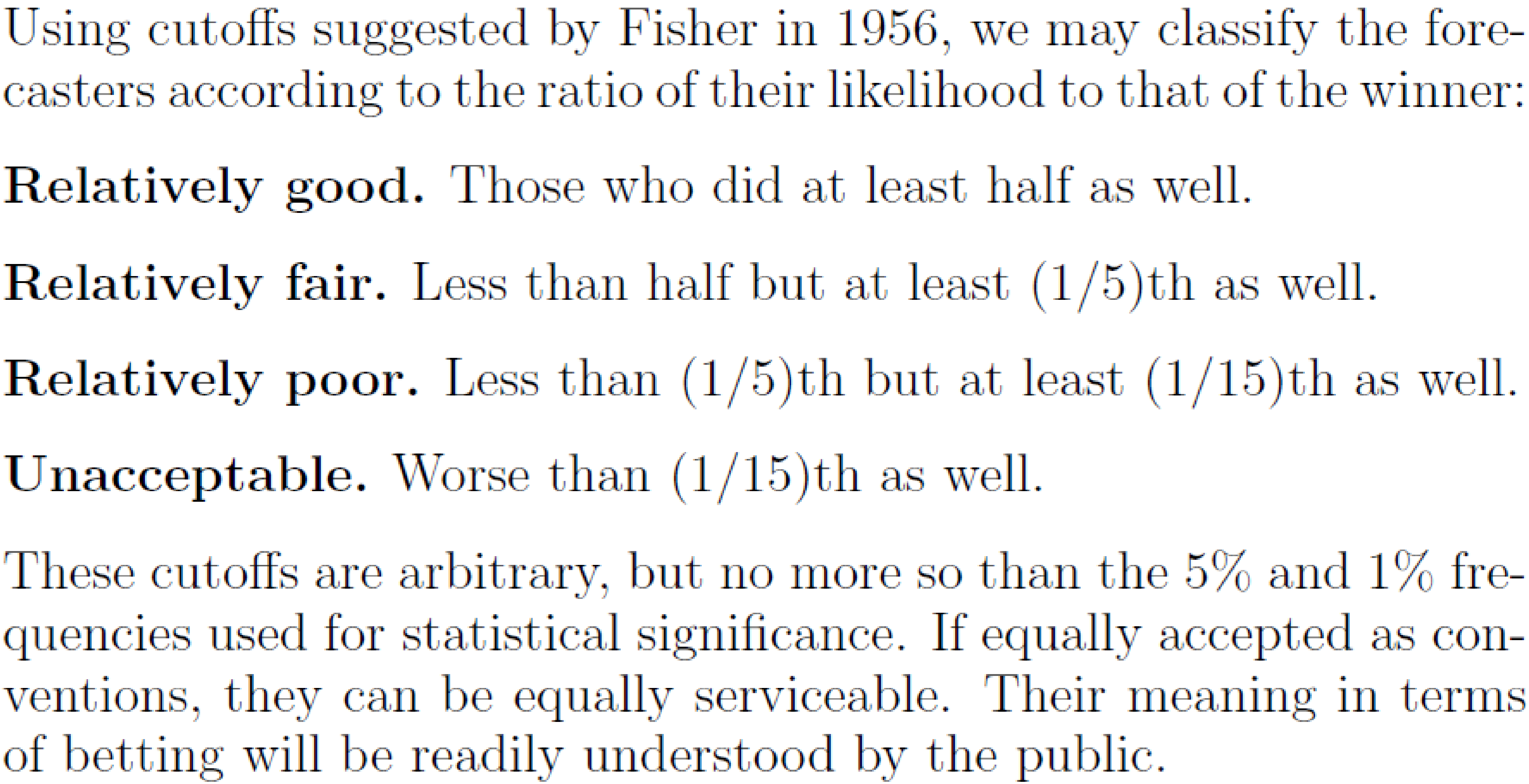
Is there any other way?

2. Description

Use model descriptively by having different parameter values bet against each other, say by Kelley betting.

Purely descriptive interpretation of relative likelihood:

- Model is conventional.
- Maximum likelihood estimate is best description within model.
- Relative likelihoods indicate the description’s precision.



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.

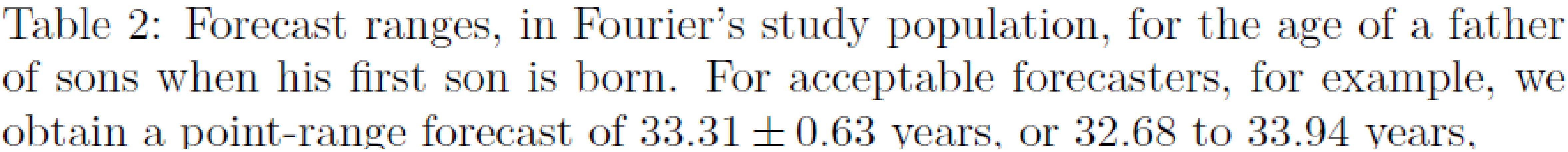
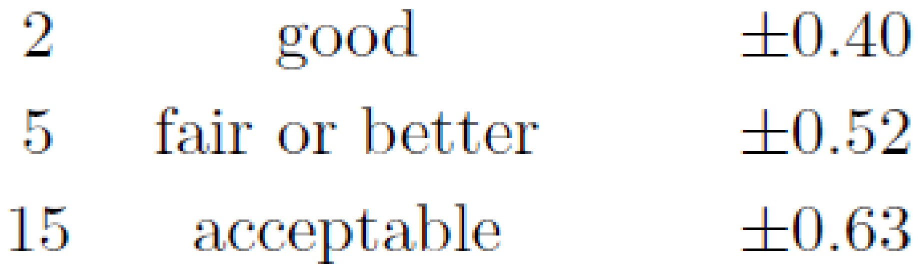


Table 2: Forecast ranges, in Fourier's study population, for the age of a father of sons when his first son is born. For acceptable forecasters, for example, we obtain a point-range forecast of 33.31 ± 0.63 years, or 32.68 to 33.94 years,



2	good	±0.40
5	fair or better	±0.52
15	acceptable	±0.63

3. Protocol

Betting makes the need for a sampling plan salient and non-ignorable.

The plan, even if it is made progressively, is more basic than the probabilities.