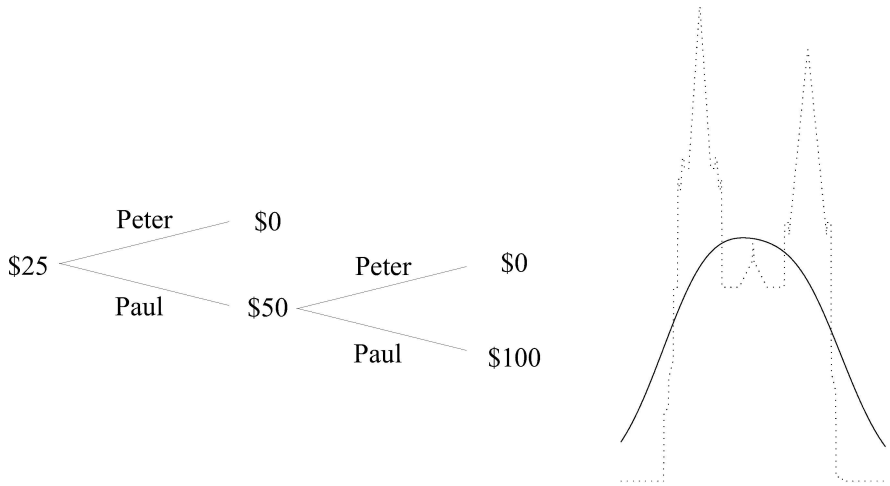
# Descriptive probability

Glenn Shafer, Rutgers University




**The Game-Theoretic Probability and Finance Project**

Working Paper #59

First posted September 29, 2021. Last revised September 30, 2021.

Project web site:
`http://www.probabilityandfinance.com`

# Abstract

Statistical modeling is often undertaken in order to make inferences from data to a larger population or to a "data-generating mechanism". But when the data is from a convenience sample or an entire population, it may be more appropriate to think of the model, with its estimated parameters, merely as a description of the data. In this case, we cannot interpret statements about the imprecision of the estimated parameters as statements of uncertainty, for there is no unseen truth to be uncertain about. Instead, we can interpret the probabilities in the models as forecasts. When different parameter values (or different algorithms) in a statistical model use Kelly betting to test each others' forecasts within the data, their relative success can be measured by a generalized likelihood. The range of forecasts made by the relatively successful algorithms, those whose generalized likelihood is a specified fraction of the maximum, can be an understandable description of the data.


# Contents

# 1 Introduction

In many branches of science, researchers routinely use significance tests and other methods of statistical inference without believing the assumptions on which they are based. As the prominent sociologist William H. Starbuck wrote in 2016 [66, p. 171],

> ... the practice of making unjustified assumptions about randomness is so prevalent that most researchers see this as conventional behavior ...

What goals lead researchers to use inferential methods in this way? Are there better ways to achieve these goals?

Some critics argue that when our study population is not a random sample from some larger population or data-generating mechanism, our only reasonable goal is description. Estimated parameters, such as regression coefficients are legitimate elements of a description, they advise, but we should dispense with significance tests and confidence intervals.[1] Researchers often resist this advice, because the techniques of statistical inference seem to help us describe data. Statistically significant differences are salient on the data landscape; perhaps we can consider differences that do not reach this threshold details. Confidence intervals for parameter values indicate, it seems, how much the description can be varied without much loss in validity.

Here we have a muddle. We are using a language and a mathematical technology that is concerned with uncertainty about unseen truths to gauge the relative validity of descriptions. This is confused and confusing.

The thesis of this paper is that we can better support the descriptive use of statistical models by interpreting them as collections of forecasting algorithms rather than as a hypotheses about the unseen.

Consider a study population in which we measure a variable $Y$ and variables $X_1, \ldots, X_K$. Suppose we want to describe this population with a family of algorithms $(P_\theta)_{\theta \in \Theta}$, where each algorithm $P_\theta$ uses the $X_k$ to give probabilities for $Y$. A reasonable first step is to evaluate and rank the $P_\theta$ according to the success of their forecasts within the study population. Then we can consider a smaller family, say $(P_\theta)_{\theta \in \Theta_{\text{Good}}}$, where $\Theta_{\text{Good}}$ consists of those $\theta$ whose forecasts performed well. The range of forecasts made by this smaller family provides one description of the study population. For some aspects of $Y$ and some values of the $X_k$, the range of forecasts may be tight enough to be interesting.

Like any method of description, this one requires relatively arbitrary choices and conventions. First we must choose $Y$ (the *target* variable) and the $X_k$ (the *forecasting* variables). Inferential statistical theory has accustomed us to think of these choices as causal modeling, but this can be pretentious and misleading. When description is our goal, $Y$ and the $X_k$ are simply what we want to describe. For whatever reason, we want to know how they vary together in the study population.

[1] See, for example, Richard Berk's "Three Cheers for Description", on pp. 206–217 of [4].

We always want to limit the complexity of the family $(P_\theta)_{\theta\in\Theta}$. In inferential statistics, simplicity is said to be a virtue of a model because complex models are likely to overfit — i.e., not to generalize to other study populations. When are goal is description rather than inference, either because we are uninterested in other populations or because we cannot make assumptions that would justify inference, a more immediate virtue of simplicity is salient. Description is description only when it is simple enough to be understood.

What family of algorithms $(P_\theta)_{\theta\in\Theta}$ do we use? What tests do we use to evaluate the performance of each $P_\theta$? What counts as performing well? Here the answers may be more conventional. Developing conventions for description is an appropriate task for theoretical statistics, and some existing statistical methods can be seen as contributions to this task.

This paper uses the method of testing probability forecasts called *game-theoretic* in [64, 65]. Here probability distributions play a dual role. On the one hand, a probability distribution is a forecast. On the other hand, it gives guidance for betting against such forecasts. When this guidance is implemented using Kelly betting, the outcome of the bet is a likelihood ratio. This suggests a way of evaluating the relative performance of the $P_\theta$: have them against each other and rank them by the resulting likelihood ratios.

The next section, §2, gives two examples to illustrate the need for descriptive probability. Then §3 develops theory of descriptive probability, §4 discusses how we can overcome the entrenched interpretation of probability as frequency and belief, and §5 summarizes the argument.

# 2 Two examples

Here are two very simple examples where the study population is not a random sample. In the first it is a convenience sample; in the second an entire population. In both examples, ostensibly inferential error probabilities awkwardly and inadequately fill a descriptive role. The theory of this paper will be applied to these examples in §3.

## 2.1 Example 1, Fourier's masculine generation

Let's begin at the beginning. The calculation of error probabilities from statistical data was first made possible by Laplace's central limit theorem, and the calculation was first explained to statisticians by Joseph Fourier (1768–1830).

Fourier had been an impassioned participant in the French revolution and an administrator under Napoleon. After the royalists regained power, a former student rescued him from impoverishment with an appointment to the Paris statistics bureau [56]. This assignment left him time to perfect the theory of heat diffusion for which he is best known, but as part of his work at the statistics bureau, he published two marvelously clear essays on the use of probability in statistics, in 1826 and 1829 [32, 33]. As Bernard Bru, Marie-France Bru, and Olivier Bienaymé have noted, these were apparently the only works on

mathematical probability read by statisticians in the early 19th century [12, p. 198].

To illustrate Laplace's asymptotic theory, Fourier studied data on births and marriages gleaned from 18th-century parish and governmental records in Paris. He was particularly interested in the length of a masculine generation — the average time, for fathers of sons, from the father's birth to the birth of his first son. On the basis of 505 cases, he estimated this average time to be 33.31 years. In Table 64 of the bureau's report for 1829 [33, pp. 143 ff], he gave bounds, in months, on the estimate's 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 |

In modern terminology, the second bound can be read as a 95% confidence interval of 33.31 years ±7.9932 months.

For Laplace's theory to be valid, the 505 cases must be mutually independent. Were the 505 fathers of sons a random sample from all 18th-century fathers of sons in Paris. Surely not. It was what the bureau could find — what has sometimes been called a *convenience sample*. So what meaning can be given to the error probabilities?

We need a different theory.

## 2.2 Example 2, a fictional survey of perceptions

Some organizations in the United States have recently surveyed their employees about perceptions of discrimination. To avoid the complexities involved in real examples, consider the following fictional example.

An organization wants to know whether its employees of different genders and racial identities differ systematically in their perception of discrimination. Most of the employees respond to a survey asking whether they have suffered discrimination because of their gender or race. The employees saying yes are distributed as shown in Table 2.2.

According to the usual test for the difference between two proportions, the difference between the rows (male vs female) and the difference between the columns (BIPOC vs White) are both statistically significant at the 5% level. But the 20 percentage-point difference between BIPOC males and BIPOC females is not, as its standard error is

$$\sqrt{\frac{1}{3}\frac{2}{3}\left(\frac{1}{20}+\frac{1}{10}\right)} \approx 0.18 = 18 \text{ percentage points.}$$

These simple significance tests seem informative. The differences declared statistically significant seem general enough to be considered features of the organization, while the difference declared not statistically significant, because it might be attributed to the idiosyncrasies of two or three people, seems less general.

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

Yet the theory of significance testing does not fit the occasion. Have the individuals in the study (or their responses to the survey) been chosen at random from some larger population? Certainly not. For anyone who has been inside an organization long enough to see its employees come and go, seeing or guessing the reasons, the idea that they constitute a random sample is phantasmagoria. Nor can we agree that their responses are independent with respect to some data-generating mechanism. Many of them see the same media and talk with each other.

If we took the theory seriously, we would also worry about multiple testing. The 5% error rate we claim for our tests is valid under the theory's assumptions only when we make just a single comparison. We have made three comparisons and might make more.

The theory's assumptions are not met, and we have abused the theory. But perhaps these are minor objections. The theory is irrelevant from the outset, because its goals are irrelevant. The organization did not undertake the survey in order to make inferences about a larger or a different population or about some data-generating mechanism. The organization wanted only to know about itself. It wanted to know how its employees' perceptions vary with gender and race. Again, we need a different theory. We need a descriptive theory.

## 2.3 R. A. Fisher's descriptive theory

Statistics began as description. Eighteenth-century statistics was description without probability, and nineteenth-century statistics was description with very little probability. Even twentieth-century statistics began with description. Karl Pearson's early work, with its emphasis on frequency curves, was unabashedly descriptive.

To his credit, R. A. Fisher tried to reconcile his small-sample theory of

statistical modeling with the descriptive tradition. In the pathbreaking 1922 paper in which he distinguished between *parameter* and *statistic* and fixed the notion of a statistical model that has endured for a century, Fisher wrote,

> ..., briefly, and in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data. This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters...

The word *regarded* is crucial here; as Fisher repeated from time to time in course of his career, the hypothetical infinite population is in the mind of the statistician. In 1925, he further explained that the statistician is only asserting that the data is "typical" of a random sample from the specified hypothetical infinite population [27, p. 701].

We see here an "as-if" mode of description: we describe the data by saying that it looks as if it were a random sample. This is coherent, and it also fits Fisher's general conception of the relation between statistical work and scientific. Scientific inference operates by induction over multiple studies and experiments, and it only it is only when we find that the same description fits many studies that we can say a theory has been confirmed.

Influential as Fisher was, his "as-if" interpretation of statistical models has not become part of the statistical tradition.[2] Its weakness is clear. So long as our culture insists on giving probability itself some reference outside the data, the vague judgment that the sample is typical of a random sample quickly drifts into an assumption about what lies outside the study population. A descriptive theory that uses probability needs a theory of descriptive probability.

# 3 Likelihood as description

We now develop the idea sketched in the introduction: interpret likelihood as a result of betting and therefore as a description of a study population.

We make likelihood descriptive by imagining that probability distributions bet against each other. In order to make the generality and simplicity of this idea clear, let us first consider the case where successive probability forecasts are made not by an algorithm but by a forecaster whose methods (if any) are not known to us. What we learn in this picture applies when probability forecasts are given by algorithms, which may or may not use information they do not forecast (forecasting variables).

[2] Notably, David J. Hand, in his recent celebration of Fisher's 1922 paper [44], does not mention it.

Because description is not inference, the descriptive use of likelihood must be distinguished from its troubled inferential use. This point is elaborated at the end of this section, in §3.6.

## 3.1 Competing probability forecasters

Suppose we observe successively $N$ variables $Y_1, \ldots, Y_N$. Suppose $M$ forecasters, who see the successive variables along with us, forecast each $Y_n$ after seeing the preceding values $y_1, \ldots, y_{n-1}$ and perhaps much other information as well. Each forecast is a probability distribution. For each $n$, we assume that the forecasters' probability distributions for $Y_n$ all have densities, discrete or continuous, with respect to the same measure, and we write $P_m^n$ for the density of Forecaster $m$'s forecast of $Y_n$. (This formulation allows the sample space to vary with $n$, but we will not use this generality in this paper.)

Perhaps we have no expertise about $Y_1, \ldots, Y_N$ ourselves. We do not know exactly how the forecasters are making their forecasts. We do not assume that there is a true or correct probability distribution for each $Y_n$ given $y_1, \ldots, y_{n-1}$, and we certainly do not know one. The forecasters themselves constitute our standard for what can be done. But after seeing their forecasts and the outcomes $y_1, \ldots, y_N$, we want to pick out those who have done a good job.

How might Forecaster 2 bet against Forecaster 1? The most natural way is *Kelly betting*. Suppose the two forecasters have just given their forecasts for $Y_n$, $P_1^n$ and $P_2^n$, and $Y_n$ is about to be observed. The ratio

$$\frac{P_2^n(Y_n)}{P_1^n(Y_n)} \tag{1}$$

has expected value 1 under $P_1^n$ and is therefore a bet against $P_1^n$. Of all bets against $P_1^n$ that would cost Forecaster 2 a unit of capital, it is the one that maximizes his own expected value of the logarithm of his subsequent capital and hence can best help him maximize his rate of growth over multiple bets.[3]

If Forecaster 2 begins by paying one unit of capital for $P_2^1(Y_1)/P_1^1(Y_1)$ and then, for each $n$ from 2 to $N$, invests all his winnings from his first $n-1$ bets in a bet proportional to (1), then his capital after the $N$ observations will be

$$\frac{P_2^1(y_1)}{P_1^1(y_1)} \cdots \frac{P_2^N(y_N)}{P_1^N(y_N)}. \tag{2}$$

Having risked one dollar, say, he walks away from the betting with the amount (2) in dollars. He may have lost money, or he may have made a great deal.

When each forecaster bets against the other in this way, the result will be a ranking, from the largest to the smallest value of

$$P_m^1(y_1) \cdots P_m^N(y_N). \tag{3}$$

[3]For more on Kelly betting, see [23, Chapter 10] and [9, 75]. For other roles Kelly betting can play in statistical theory, see [64, 73]. An important alternative to Kelly is fractional Kelly betting, which risks only a fixed proportion of current capital on each bet. Being more cautious, this penalizes the forecasters less for extreme errors. This potential robustness merits study, but it is outside the scope of this paper.

Generalizing Fisher, let us call (3) Forecaster $m$'s *likelihood*.

Our culture tells us how to think about betting results. The forecaster with the largest likelihood is the winner of the contest, perhaps not the very best among the forecasters, but lucky enough this time and entitled to claim a place among the best. As in any contest that combines skill and luck, our assessment will depend on what we know about the credentials of the competitors and the difficulty and extent of the task. But any competitor who falls far short has been discredited. If Forecaster $m$ has the highest likelihood and Forecaster $m'$'s likelihood is only (1/15)th as great, then Forecaster $m'$'s competence will, in Fisher's words, be open to grave suspicion. Forecaster $m$, betting against him, has risked \$1 and come away with \$15.

Using cutoffs suggested by Fisher in 1956 [30, p. 71],[4] 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.

## 3.2 Competing probability forecasting algorithms

Fisher was not writing about probability forecasters. He was writing about a statistical model — a class of probability distributions $(P_\theta)_{\theta\in\Theta}$. But the $P_\theta$ can be thought of as probability forecasters, and the preceding discussion applies; their likelihood ratios can be interpreted as the factors by which each has multiplied its money betting against the other. This is equally true for any collection $(P_\theta)_{\theta\in\Theta}$ of algorithms that make probability forecasts.

Let us call a collection $(P_\theta)_{\theta\in\Theta}$ of such algorithms a *forecasting family*. Just as Fisher, in his celebrated 1922 paper on theoretical statistics [26, p. 314], treated the selection of a statistical model as an empirical matter, to be left to the "practical statistician", we expect the statistician to choose the forecasting family after studying the data.

We call the elements of $\Theta$ *forecasters*. Each $\theta \in \Theta$ forecasts variables $Y_1, \dots, Y_N$, Forecaster $\theta$'s forecast $P_\theta^n$ being a probability distribution for $Y_n$. It is convenient to say that the forecasters all have the same information available, although some may ignore some of this information. Similarly, it is convenient to say that the forecasters observe the $Y_n$ in order, meaning that $y_1, \dots, y_{n-1}$ are available for forecasting $Y_n$, but some or all of the forecasters may ignore

[4] The suggestion appears on page 75 of the posthumous third edition (1973).

this information as well. The number $N$ may be equal to one. We call the collection of $N$ individuals for which $Y$ and the other information is collected the *study population*.

We call the product

$$P_\theta^1(y_1)\cdots P_\theta^N(y_N) \tag{4}$$

Forecaster $\theta$'s *likelihood*. We assume the forecasting family $(P_\theta)_{\theta\in\Theta}$ has been chosen so that there is always a forecaster, say $\hat{\theta}$, who has the greatest likelihood, and we call the ratio

$$L(\theta) := \frac{P_\theta^1(y_1)\cdots P_\theta^N(y_N)}{P_{\hat{\theta}}^1(y_1)\cdots P_{\hat{\theta}}^N(y_N)} \tag{5}$$

Forecaster $\theta$'s *relative likelihood*. The natural logarithm, $\ln(L(\theta))$ denoted $l(\theta)$, is the *log relative likelihood*.

The information used in making the forecasts is left implicit in (4) and (5). It is also implicit in our terminology. We call (4) simply $\theta$'s likelihood, not its conditional likelihood given the information used, as in inferential theory [60, p. 155].

To use the relative likelihoods descriptively, we consider the $\theta$ for which $L(\theta)$ is relatively good (at least $1/2$) and describe the ranges of forecasts they make. We might similarly describe the range of forecasts for which $L(\theta)$ is acceptable (at least $1/15$). The virtues of these ways of describing the data include their honesty and transparency. They are honest because they do not use assumptions we do not believe. They are transparent because their methods of assessing the forecasting algorithms can be explained in terms of betting even to those who have not studied statistics.

A probability distribution is generally a complex object. So summaries are needed. When $Y$ takes numerical values, the expected value is often a useful summary; we may call it the *point forecast*. For particularly interesting values of the forecasting variables, we may choose to report ranges of point forecasts for the good performing or acceptably performing $\theta$. We may call these ranges *point-forecast ranges*. We may also be interested in a range of values for $Y$ to which all the good performing $\theta$ assign high probability. We might report the union of all the interquartile ranges predicted by good performing $\theta$, or perhaps the union over the good performing $\theta$ of some other interval prediction. These may be called *interval-forecast ranges*. What ranges the statistician reports will depend on what she and her audience want to know about the study population. When we want to know what is typical of the study population, the point-forecast range is appropriate. An interval-forecast range, if it disagrees sharply with the empirical spread of the data, may lead us to question our choice of the forecasting family. In general, point-forecast ranges may add more than interval-forecast ranges to empirical variances, quantiles, and other descriptive statistics available directly from the data. Here we may want to take a lesson from the French mathematicians of the 19th century, whose sophisticated Laplacean statistical theory was displaced in geodesy by a Gaussian methodology less interested in inference than in best estimates [46, 62].

How do we choose a forecasting family? As in other domains of description, it will be useful to have conventional choices, and so we will prefer to use families that are already well known. The choice of a family includes the choice of forecasting variables (when forecasting is by multiple regression, for example), and here we begin with the questions we are asking. In Example 1, for example, the organization was asking about how perception varies by minority status and gender, not how it varies by age, health status, or other features of the organization's employees.

In both our examples, Example 1 and Example 2, the target variable and the forecasting variables of interest were specified before the data were gathered. But in other cases, a statistician may come to data uncertain about what target and forecasting variables might permit interesting descriptions. Identifying them requires exploring the data – *exploratory data analysis*, John Tukey called it.

Forecasting within a study population generally looks better and better as we increase the number of forecasting variables. In inferential statistics, we learn that this can lead to overfitting, giving a misleading picture of the target superpopulation. This fact is not irrelevant in cases where we want to compare descriptions of different but similar study populations. But when description is the goal, simplicity also has a more direct value. The most complete description of data from a study population is the data itself. Useful description requires the reduction of data.

In addition to having the forecasters in the family test each other by betting, the statistician might want to implement a betting strategy to test the forecasting adequacy of the entire family. This is surely advisable if the statistician is concerned about a particular way the family might fail and has at hand a corresponding alternative family. On the other hand, experience suggests that traditional comprehensive goodness-of-fit tests of conventional families can be undiscerning [10]. When observations are made sequentially, machine-learning tests of randomness have the opposite but equal problem: they generally reject [72]. So we should always bear in mind that the choice of a family is a convention made for the purposes of description, not an empirical assumption.

We can, of course, later use our relatively good algorithms to forecast outside of the study population. This is inference in only a weak sense. We are trying outside the study population what worked inside. If a particular forecasting family is repeatedly successful when used in this way, then we engaged in induction and inference. We can think of this as a shift from exploratory data analysis, where we are looking for good descriptions, to confirmatory data analysis, where we are looking for descriptions so stable over time and circumstance that they may be given causal meaning.

Let us now apply this paper's proposal to three simple examples. The first is the simplest possible example, where a single probability is used to forecast an event for each individual in a study population. The other two are the examples we considered in §2.

As we will see, the proposal requires much more computation than the statistical analyses it might replace. This will be even more true when we undertake

to apply the proposal to multiple regression and other statistical models where it might be most useful.

### 3.3 Example 0, forecasting with a single probability

Suppose we observe successive trials of an event, and each algorithm in our forecasting family has a fixed probability that it uses each time as its forecast. Formally, $\Theta = (0, 1)$, and Forecaster $\theta$ always gives $\theta$ as the probability that the event will happen.

If we observe 100 trials, and the event happens 70 times, then

$$L(\theta) := \left(\frac{\theta}{0.7}\right)^{70} \left(\frac{1-\theta}{0.3}\right)^{30}.$$

Our scheme for rating the forecasters yields these point-forecast ranges:

**Relatively good:** $L(\theta) > \frac{1}{2}$, or $0.64 < \theta < 0.76$.

**Fair or better:** $L(\theta) > \frac{1}{5}$, or $0.61 < \theta < 0.78$.

**Acceptable:** $L(\theta) > \frac{1}{15}$, or $0.59 < 0.80$.

The forecaster $\theta = 1/2$ may have been of particular interest, and we may want to emphasize that its performance was unacceptable.

Not surprisingly, Fisher's categories are consistent with inferential practice. The standard error of the maximum-likelihood forecaster 0.7 is 0.046, suggesting a 95% confidence interval of $(0.61, 0.79)$, very significantly different from $1/2$. But unlike the analogous inferential model and significant tests, the forecasting family and the resulting data analysis merely describe the data. The forecasting family does not say that the trials of the event are in any sense independent. The analysis tells us merely which constant probabilities perform relatively well in the data.

### 3.4 Example 1, Fourier's masculine generation, continued

Recall Fourier's estimation of the average age in 18th-century French fathers of sons when their first son was born. From the ages of 505 fathers, say $y_1, \ldots, y_{505}$, he found the empirical average

$$\overline{y} := \frac{\sum_{n=1}^{505} y_n}{505} = 33.31 \text{ years}, \tag{6}$$

and the empirical standard deviation[5]

$$s := \sqrt{\frac{\sum_{n=1}^{505} (y_n - \overline{y})^2}{505}} = 7.642 \text{ years}. \tag{7}$$

[5]The empirical standard deviation given here is calculated from the error limits Fourier gave. He did not report the data, and we know that he worked not with standard deviations and standard errors but with what was later called the *modulus* by some authors, equal to $\sqrt{2}$ times the estimate's standard error.

Using modern language, we may say that Fourier assumed that the 505 ages were independent random variables, treated 33.31 as an estimate of their common mean, and used the central limit theorem to obtain approximate error limits on this estimate.

For a descriptive analysis, we do not need Fourier's assumptions. We need a forecasting family. Let's use the most familiar one, the normal family with mean $\mu$ and variance $\sigma^2$. Here $\theta = (\mu, \sigma^2)$, $\mu$ is Forecaster $\theta$'s point forecast of each $Y_n$, and $(\overline{y}, s)$ is the maximum-likelihood forecaster. The log relative likelihood for Forecaster $(\mu, \sigma^2)$ is

$$l(\mu, \sigma^2) = N\left(\ln(s) - \ln(\sigma) - \frac{s^2 + (\overline{y} - \mu)^2}{2\sigma^2} + \frac{1}{2}\right),$$

where $N = 505$. For fixed $\mu$, this is maximized by setting $\sigma^2$ equal to $s^2 + (\overline{y} - \mu)^2$.[6] So to see the best we can do with the point forecast $\mu$, we consider the log relative likelihood

$$l(\mu, s^2 + (\overline{y} - \mu)^2) = N\left(\ln(s) - \frac{1}{2}\ln(s^2 + (\overline{y} - \mu)^2)\right) \tag{8}$$

This is greater than $\ln(1/C)$ when $\mu$ is in the interval

$$\overline{y} \pm s\sqrt{(2C)^{\frac{1}{N}} - 1}. \tag{9}$$

Table 2 uses (9) to calculate good, at least fair, and acceptable point-forecast ranges. Associating each point forecast $\mu$ with the 95% probability forecast

$$\mu \pm 1.96\sqrt{s^2 + (\overline{y} - \mu)^2},$$

we also obtain the interval-forecast range given in the last column. The point-forecast ranges in the table can be compared with Fourier's 95% range of 33.31 years $\pm$7.99 months or 0.666 years.

Perhaps other information in the data gathered by the Paris statistics bureau could have allowed better forecasts. Perhaps, for example, there was a discernible trend from the beginning to the end of the 18th century. But the question we address here — perhaps also the question Fourier was really asking — is which constant forecasts do a good job.

## 3.5 Example 2, fictional survey, continued

Here people were asked a yes-no question, and so a probability forecast is a single number. But now a forecaster has information on which to base the probability — which of the four groups the individual belongs to. So a forecaster is defined by four probabilities:

$$\theta = (p_{\mathrm{bf}}, p_{\mathrm{bm}}, p_{\mathrm{wf}}, p_{\mathrm{wm}}),$$

[6] In inferential theory, the result of maximizing a likelihood over an unwanted parameter is sometimes called a "profile likelihood" [60, p. 158].

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 $33.31 \pm 0.63$ years, or 32.68 to 33.94 years, and a 95%-interval-forecast range of $33.31 \pm 10.02$ years, or 23.29 to 43.33 years.

| | | ranges (intervals around 33.31) | |
|---|---|---|---|
| $K$ | class | point-forecast | interval-forecast |
| 2 | good | ±0.40 | ±6.40 |
| 5 | fair or better | ±0.52 | ±8.26 |
| 15 | acceptable | ±0.63 | ±10.02 |

where $p_{\mathrm{bf}}$ is the forecast that a BIPOC female will say yes to the survey, etc. According to the data in Table 2.2, the maximum-likelihood forecaster is

$$\hat{\theta} = \left( \frac{8}{10}, \frac{12}{20}, \frac{20}{50}, \frac{20}{120} \right),$$

and the relative likelihood is

$$L(\theta) = \left( \frac{p_{\mathrm{bf}}}{4/5} \right)^{8} \left( \frac{(1-p_{\mathrm{bf}})}{1/5} \right)^{2} \left( \frac{p_{\mathrm{bm}}}{3/5} \right)^{12} \left( \frac{(1-p_{\mathrm{bm}})}{2/5} \right)^{8} \left( \frac{p_{\mathrm{wf}}}{2/5} \right)^{20} \left( \frac{(1-p_{\mathrm{wf}})}{3/5} \right)^{30} \left( \frac{p_{\mathrm{wm}}}{1/6} \right)^{20} \left( \frac{(1-p_{\mathrm{wm}})}{5/6} \right)^{100}.$$

We found earlier that the 20 percentage-point difference between BIPOC males and BIPOC females is not statistically significant. In our theory of descriptive probability, the question can be reframed this way: what differences between BIPOC males and BIPOC females within the study population are forecast by good forecasters? We can answer the question by looking at all the $\theta = (p_{\mathrm{bf}}, p_{\mathrm{bm}}, p_{\mathrm{wf}}, p_{\mathrm{wm}})$ that rank as good forecasters by having a value of $L(\theta)$ greater than 1/2 and finding the range of their values for $p_{\mathrm{bf}} - p_{\mathrm{bm}}$. The range is from a little more than 0 up to about 0.4.

When the individuals responding to a yes-no survey are categorized in more than one way, or when other data is collected about them, we may prefer to use a more sophisticated forecasting family, such as logistic regression. The logic will remain the same. For particular interesting values of the forecasting variables, we can give the range of probability forecasts given by good forecasters, and we can similarly give ranges for differences in probabilities or for odds ratios. The computations involved are not trivial, but software environments adequate to the task would not be need to be more complex than those that now use logistic regression for nominally inferential analyses.

The descriptive approach can be compared with the inferential approach used in 2016 by the University of Michigan's Diversity, Equity & Inclusion Initiative.

Michigan sought inferential legitimacy by surveying random samples. As they explained in their report on the faculty survey [54, p. 6],

> Given the large faculty population at the University of Michigan, this study used a sample survey approach rather than a census of all faculty. A carefully selected sample, with randomization, allows researchers to make scientifically based inferences to the population as a whole.

In the case of the faculty, 1,500 out of 6,700 faculty members where chosen at random to complete the survey. The survey results were then analyzed using logistic regression, and a number of differences were observed to be statistically significant. It was found, for example, that female faculty were 130% more likely to feel discriminated against than male faculty (i.e., the odds ratio for a positive response to the question was 2.3 and significantly different from 1).

The results of the survey were clearly meaningful, but the inferential logic is problematic. As David A. Freedman has shown, randomization probabilities do not justify logistic regression [35]. Our descriptive theory is not affected by this problem and is just as applicable to a complete census as to a random sample.

### 3.6 Inferential likelihood

The intuition that likelihood has inferential significance goes back at least to the late 18th century [68]. The postulate that the happening of an event supports competing causes in proportion to the probability they give to the event was about all Laplace offered to justify his first formulation of Bayes's rule [67]. Fisher pried the postulate out of its Bayesian frame, coined the name *likelihood*, and eventually suggested that the numerical ranking of rival probability distributions by their likelihood can stand on its own as a report on statistical evidence. Some statisticians have found Fisher's version of the postulate appealing. A. W. F. Edwards and Richard Royall wrote whole books elaborating it [22, 60], giving many intuitively reasonable examples. But when applied generally, it can encounter fatal difficulties, which are summarized concisely by David R. Cox and David V. Hinkley in their well known textbook [18, pp. 50–52].

One of Cox and Hinkley's examples, particularly relevant here, calls to mind a lottery. We will observe an integer from the set $\{1, \ldots, 100\}$, and we consider 101 probability distributions for this observation, indexed by $\Theta := \{0, 1, \ldots, 100\}$. The distribution $P_0$ assigns equal probabilities to the 100 possibilities, while for $k$ between 1 and 100, $P_k$ gives probability 1 to $k$. When the observation turns out to be $x$, say, $\theta = x'$ has likelihood zero if $x'$ is not equal to $x$ or 0, and the likelihood of $\theta = x$ is 100 times that of $\theta = 0$. Interpreting this ratio as evidential support is unsatisfactory because, as Cox and Hinkley put it, "even if in fact $\theta = 0$, we are certain to find evidence apparently pointing strongly against $\theta = 0$."

What does this example tell us about the descriptive use of likelihood? The best answer, perhaps, is that in this case the likelihood is a valid but uninter-

esting description. There is no reduction of the data. We are simply told that whoever won the lottery won the lottery.

# 4 Probabilities as forecasts

This paper's proposal faces two serious conceptual hurdles. The first is the mindset that insists on interpreting probabilities as hypotheses about a hidden reality. The second, more specific and perhaps even more entrenched, is the notion that probabilities are ultimately and fundamentally frequencies.

As I now argue, in §4.1, the first of these hurdles is most effectively addressed by history. "To penetrate to the reasons of things," Marie-France Bru and Bernard Bru have advised, "look at how they have gradually been revealed in the course of time, in their progressions and in their ruptures..." [13, pp. 301–302]. Significance testing is not a method that recently strayed after being correctly used for centuries. It was troubled from the beginning. Recognizing this will help legitimize trying something different.

As I argue in §4.2, escaping from the more specific muddle of frequentism is another matter. For more than a half century, many prominent statisticians have campaigned to replace frequency with belief. But while this campaign has readied us to question the identification of probability with frequency, belief does not provide a foundation for description. As I have just suggested, forecasting does provide such a foundation. In order to overcome the hurdle of frequentism we need a theory of probability in which frequency is replaced by forecasting. This theory is provided by the game-theoretic framework developed in [65, 64], in which probability forecasts are taken as hypothetical betting offers and tested by algorithms that pretend to make some of the offered bets and discredit the forecaster to the extent that their nominal bets succeed.

## 4.1 The enduring fiction of randomness

Research and teaching in mathematical statistics emphasizes independent and identically distributed observations from probability distributions — i.e., random samples. Yet in applications non-random samples are the rule, not the exception. So it has always been. We have been so successful in closing our eyes to this enduring contradiction that we must excavate its history before we can fully appreciate the legitimacy of descriptive probability.

**Eighteenth-century beginnings.**

In 1703, when Jacob Bernoulli asked Gottfried Wilhelm Leibniz for help in finding data on human mortality to which he could apply his celebrated theorem, the first law of large numbers, Leibniz was quick to point out that such data could not possibly satisfy the assumption of stability the theorem required:

> The difficulty in it seems to me to be that contingent things or things that depend on infinitely many circumstances cannot be determined

> by finitely many results, for nature has its habits, following from the return of causes, but only for the most part. Who is to say that the following result will not diverge somewhat from the law of all the preceding ones — because of the mutability of things? New diseases attack humankind. Therefore even if you have observed the results for any number of deaths, you have not therefore set limits on the nature of things so that they could not vary in the future.[7]

Bernoulli may have been troubled by this objection. But the posthumous book containing his theorem and its proof, *Ars conjectandi*, concluded with a declaration of faith in the stability of causes:

> ... if the observations of all events were continued for the whole of eternity (with the probability finally transformed into perfect certainty) then everything in the world would be observed to happen in fixed ratios and with a constant law of alternation.[8]

Thanks to his friend Edmund Halley, Abraham De Moivre had the kind of mortality data that Bernoulli had sought. Beginning in the 1720s, he used it profitably, fitting a mostly linear model and using this model to help British aristocrats price life-time leases of their land to farmers [3]. It is doubtful that he believed that Halley's data on 17th century-mortality in Breslau applied precisely to 18th-century British farmers. But in the 1730s, when he proved the central limit theorem for binomial observations, he seconded Bernoulli's vision of eternal sampling:

> ... altho' Chance produces Irregularities, still the Odds will be infinitely great, that in process of Time, those Irregularities will bear no proportion to the recurrency of that Order which naturally results from original Design. [19, p. 243 of 2nd edition, p. 251 of 3rd edition]

Bernoulli had imagined sampling with replacement from an urn containing a finite number of black and white tokens in unknown proportion. Pierre Simon Laplace, when introducing his version of Bayesian inference in 1774, imagined sampling without replacement from an urn with "an infinity of white and black tickets in unknown ratio".[9] This sampling metaphor has been familiar to statisticians ever since. But it has seldom described the samples we actually use.

**Nineteenth-century failure in France.**

The general version of the central limit theorem that Laplace derived in 1810 permitted the calculation of error probabilities for averages and proportions.[10]

[7] Translation by Edith Sylla [8, p. 39].

[8] Translation by Edith Sylla [8, p. 339].

[9] Translation by Stephen Stigler [67, p. 365]. See also Laplace's letter to Georges Louis Le Sage in 1784 [43, vol. I, letter 72], discussed by Marie-France and Bernard Bru [13, vol. 1, p. 96].

[10] De Moivre had earlier derived the normal approximation for binomial probabilities. For a comprehensive history of the central limit theorem, see [25].

As we have noted, statisticians learned the method from Joseph Fourier in the 1820s. Its application, especially to medicine, was widely debated in France beginning in the 1830s.

When explaining the formulas for error probabilities in his influential essays in 1826 and 1829, Fourier explained that the main underlying principle was that

> in an immense number of observations, the multiplicity of the chances makes what is accidental and random disappear, and only the sure effect of constant causes remains, so that there is no chance at all in natural facts taken in a very large number.[11]

But the method could err:

> The most common sources of error and uncertainty in the conclusions that many writers deduce from statistical studies are (1) the inexactness of observations gathered by extremely varied and incomparable methods, (2) too few observations, which prevents their division into separate series and the calculation of the result of each series, and (3) progressive or irregular changes in the causes over the period of the observations.[12]

The second point in the list relates to his recommendation of a form of cross-validation: check the consistency of data by making the same calculation on different parts of it. The second and third together were his way of giving empirical content to the independence or randomness assumption.

A favorite topic for statisticians in this period was variation in the ratio of male to female births. In 1824, the prominent mathematician Siméon-Dénis Poisson noted that the ratio was smaller for illegitimate than for legitimate births [57]. In 1830, he applied Laplace's theory to decide whether this and other variations in the ratio could have happened by chance [58]. He concluded that the difference was real, and he also found that the ratio was smaller in Paris than in the rest of France, for both legitimate and illegitimate births.

Applications in medicine were obviously more important. In 1837 the French Academy of Sciences debated whether observational data could be used to decide between two methods for removing gallstones, and the physician Jules Gavarret subsequently published a book on the use of Laplace's method to test the difference between two proportions using such data [38, 45]. J. Rosser Matthews has traced the debate in medicine inspired by Gavarett's book, noting its more positive reception in Germany [51].

It is not hard to see flaws in these early applications. We have already seen the difficulties with Fourier's prime example, the length of the masculine generation. Poisson's search for significant differences in the birth ratio looks like p-hacking, as Augustin Antoine Cournot insinuated in 1843, after Poisson's death [16]. It is very unlikely that the two treatments for gallstones were administered to similar groups of people.

[11] My translation from [32, p. x].
[12] My translation from [32, p. xv].

Cournot's book was a remarkably lucid and discerning presentation of Laplace's theory. The book remained the high point of mathematical statistics in France in the nineteenth century for an unhappy reason: neither mathematicians nor the wider intellectual public found the theory's applications very convincing. The leading mid-century mathematical statistician, Jules-Irénée Bienaymé, put most of his intellectual effort into discouraging erroneous applications [12]. By the end of the century, a flood of statistics had transformed thinking about science, medicine, and society, but probability theory had been left behind. Leading French mathematicians saw Laplace's theory as derisory. Lucien Le Cam, a 20th-century proponent of the central limit theorem was so dismayed by this attitude that he once called the theorem's most prominent detractors "the loathsome Bertrand and Poincaré" [47, p. 96].[13]

**Twentieth-century recapitulation.**

Laplace's infinite urn still sometimes appears in textbooks, but in the 20th century it was largely replaced by other metaphors. Three prominent ones are *infinite hypothetical population*, *superpopulation*, and *data-generating mechanism*

Fisher popularized the term *infinite hypothetical population* beginning in the 1920s, using it interchangeably with *hypothetical infinite population*. These words now seem old-fashioned and perhaps even naive. The meaning they give to "infinite" harks back to a concept of potential infinity that already puzzled some mathematicians in the 1920s [1]. But as we have already noted, Fisher thought about what he meant by it. Whereas some later statisticians have wanted to see the infinite population as something real, Fisher always emphasized its hypothetical nature, seeing it as invented by the statistician to describe his uncertainty.

Fisher's formulation leaves space for the statistician to judge that a sample is not random, perhaps because it represents an entire population or perhaps it is a *convenience sample*, collected in a somewhat systematic but unmodelled way rather than randomly.[14] Authors in the social science literature in the mid-twentieth century sometimes declined to use significance tests for such samples. According to Google's Ngram, *convenience sample* increased in popularity until about 1990, then declined sharply. Abstinence from significance testing also declined. In a study of the use of significance testing in two prominent sociology journals from 1935 to 2000 [48], Erin Leahey found a shift around 1975 among authors who had data on an entire population. Before that date, some of these authors declined to use significance tests, afterwards few did.

The term *superpopulation* has been used in a number of contexts. Beginning at least with Fisher in the 1930s, it has been used in reference to Bayesian

[13]The importance of statistics in the 19th century has been emphasized by Ian Hacking and other historians [42]. For details on the negative attitude towards Laplace at the end of the 19th century see [13].

[14]The term *convenience sample* seems to have appeared around 1960. In 1956 [53], John Neter contrasted random selection with "selection by convenience" and called the result a "convenient sample". Santo Camilleri commented negatively on convenience samples, with "convenience" in quotes, in 1962 [14]. The term was used less self-consciously by 1964 [70].

priors [28, 29, 71, 40]. Beginning at least in the 1940s, it has been used in industrial and agricultural contexts. In 1950 [11], Irwin Bross suggested that groups in an analysis of variance might be considered a sample from a superpopulation. In 1946 [15], William G. Cochran suggested that when sampling from a finite population in time or space, correlations between adjacent units might be modeled by thinking of the finite population as itself sampled from an infinite population with those correlations. Although Cochran did not use the term *superpopulation*, his work is sometimes seen as the beginning of the standard superpopulation setup of sampling theory.

Google's Ngram indicates that the use of *superpopulation* rose sharply from about 1970 to 1990 and remained stable thereafter. The popularity of Fisher's *infinite hypothetical population* and *hypothetical infinite population*, on the other hand, peaked around 1960 and declined sharply thereafter. They have been largely replaced by *data-generating mechanism*, which began to become popular in the 1960s.[15]

In 1976 [59], Richard Royall used *superpopulation model* and *prediction model* as synonyms. The implication is that the study population from which we estimate parameters is sampled from a larger population, and that the result will be used to predict further observations from the larger population. This study population may be a single observation. As in Cochran's original example, the modeling can then account for correlations over time or space, which are predicted to recur in the further observations. But when we see the term *data-generating mechanism* instead of *superpopulation*, there is often neither deliberate sampling nor any attempt to account for correlations. We may be dealing either with an entire population that leaves nothing to predict or with a convenience sample from which predictions are dubious.

In 1995, Richard Berk, Bruce Western, and Robert E. Weiss listed examples of entire populations that are often studied by political scientists, sociologists, and economists: all industrialized nations, all large cities in the United States, etc. [6]. Studies of such *apparent populations*, as Berk, Western, and Weiss called them, are predominant in some branches of social science, including finance and accounting. These studies usually claim to make inferences about data-generating mechanisms but say little or nothing about whether and how purported mechanisms can be used to make future predictions. Even in epidemiology, it is common for statistical studies of entire populations of certain countries or regions to be silent about their generalizability [2]. Abstinence from prediction is also the norm when data-generating mechanisms are inferred from convenience samples [5]. See also the critiques by Freedman [34] and Berk [4].

[15] The earliest use of *data-generating mechanism* in connection with a statistical model that I have seen is in a 1963 article on the cost of ship repair, which first identified the mechanism as "accounting procedures and budgetary pressures" and then identified it with a statistical model [24, p. 336]. The term appears in econometrics in the late 1960s [39, 55].

## 4.2 Replacing frequency with betting success

At one time, the term *frequency theory of probability* referred to Richard von Mises's proposal for axiomatizing the notion of a random sequence in terms of limiting frequency. When Ernest Nagel coined the term *frequentist* in 1936 [52], he was still discussing von Mises's proposal. But once Andrei Kolmogorov's axiomatization became the starting point, probability's identification with frequency has come to look like a misunderstanding of the law of large numbers. As argued here, a close at the law of large numbers shows that the empirical meaning of mathematical probability rests on high-probability predictions, which can be interpreted in terms of betting.

### Frequency or high probability?

A 95%-confidence set, we know, is a method for computing a set. We fix a class $(P_\theta)_{\theta\in\Theta}$ of probability distributions, assume that there is some $\theta \in \Theta$ such that $P_\theta$ forecasts a certain phenomenon $Y$ well, and set out to learn from the observation of $Y$ which $\theta$ this might be. We call a random subset $C(Y)$ of $\Theta$ a 95% confidence set if

$$P_\theta(\theta \in C(Y)) \geq 0.95 \tag{10}$$

for all $\theta \in \Theta$.

I rehearse this not as a prelude to rehashing the pros and cons of confidence sets but to set the stage for a not-so-innocent question: Why is a confidence set called frequentist? Where is the frequency?

Every student of statistics knows the standard answer: Were we to repeat the experiment many times, the set $C(Y)$ would contain the true $\theta$ at least 95% of the time. This is the teacher's way of explaining the meaning and importance of the probability 0.95 in (10). Frequency is thus the meaning of probability.

Really? The teacher relies, it seems, on the law of large numbers: In many repeated trials of an event with probability 0.95, *there is a high probability* that the event will happen about 95% of the time. But this is brazenly circular. It explains the meaning of high probability by means of a high probability.

Fisher, often himself called a frequentist by later commentators, was scathing about the circularity. He called it "a perpetual regression defining probabilities in terms of probabilities in terms of probabilities" [31, p. 266]. Countless other luminaries, including the probabilists Kolmogorov and Joseph Doob and the statisticians Abraham Wald and Charles Stein, have made similar points. Stein commented as follows on what Kolmogorov had written in 1933 about the empirical interpretation of probability [20, p. 460]:

> In his book he [Kolmogorov] mentions briefly two aspects of the interpretation. The first is the traditional relative frequency of occurrence in the long run. And the second is that when one puts forward a probabilistic model that is to be taken completely seriously for a real world phenomenon, then one is asserting in principle that any single specified event having very small probability will not occur.

> This, of course, combined with the law of large numbers, weak or strong, really is a broader interpretation than the frequency notion. So, in fact, the frequency interpretation in that sense is redundant.

Doob explained the role of small probabilities this way [21, pp. 201–202]:

> If one starts with mathematical probability theory the obvious general operational translation principle is that one should ignore real events that have small probabilities. How small is "small" depends on the context, for example, the demands of a client on a statistician. Somewhat more precisely, one first makes a judgment on the possibility of the application of probability in a given context; if so, one then sets up a model and comes to operational decisions based on the principle that hypotheses must be reexamined if they ascribe small probability to a key event that actually happens.

Even though countless authorities in probability and statistics have explained that the meaning of a probability model lies in its prediction that certain events of high probability will happen, or equivalently in its prediction that certain events of small probability will not happen,[16] frequency has remained central to how statisticians explain themselves to their public.

Why? Probably because it goes down easily with the uninitiated. The equation between probability and frequency has been part of our culture since the 19th century. Cournot stated so clearly that probability connects with phenomena only when it predicts the failure of an event with small probability that this principle has been called *Cournot's principle* [61]. But even he underlined the probability's identification with frequency when explaining the meaning of a confidence probability [16, §107].

Cournot's principle is difficult to communicate because it inevitably provokes a seemingly conclusive objection: Doesn't what happens always have a small probability? There are answers. The statistician specifies her events of small probability (her tests and confidence intervals) in advance. The test events are simple events. The events of small probability that actually happen are not known in advance and are too precise to be simple. Anyway, this is the way applied statistics works, like it or not. Good answers, but not good enough to forestall endless debate.

### Small probability or high betting score?

Betting is just as anchored in our culture as frequency, and equally connected with probability. Within probability theory, betting's historical credentials are even better than frequency's, for betting provided the underpinning for probability mathematics long before Jacob Bernoulli formulated his law of large numbers. It is easy to see, moreover, how bets can replace small probabilities in statistical inference.

[16] For additional examples, see [41, 49, 61, 69, 74].

You announce in advance that a particular event $E$ has a small probability 5% and so will not happen. It happens. Why is this surprising, and why might it discredit your judgement? Because when you made the announcement, I might have insisted on betting with you at the odds your probability implied, risking \$1 on $E$ and walking away with \$20. Betting against a probability is the best way to discredit it, and weathering such bets is the best way for a probability or a probability forecaster to gain credit.

As Christiaan Huygens made clear in his 17th-century introduction to the calculus of chances, a bet need not be all-or-nothing;[17] if you announce a probability distribution $P$, I might pay \$1 for any non-negative payoff $S$ to which $P$ gives expected value 1. The actual payoff, say $\$S(y)$, then discredits $P$ to the extent that it is large — i.e., to the extent that I have multiplied the money I risked by a large factor.

Markov's inequality says that

$$P\left(S(Y) \geq \frac{1}{\alpha}\right) \leq \alpha,$$

and so you might explain the force of a large betting score $S(y)$ by saying that it had small probability. But this is wrong-headed. Historically and in our wider culture, a large betting score is more fundamental than a small probability, and trying to explain it in terms of a small probability goes backward, introducing complication and confusion. If $S(y) = 30$, what is $\alpha$? For a small probability $\alpha$ to have force, it must be announced in advance along with $S$, before I know that $S(y) = 30$. But if I announced $\alpha = 0.05$, say, this would reduce my 30 to 20. There is no need to do this. The number 30 is honestly come by — a fair measure of the extent to which I have discredited your $P$.

# 5 Discussion

As a flurry of recent theoretical studies have shown, testing by betting can play an important role in inferential statistics. Because crucial assumptions are so often unjustified in nominal applications of inferential statistics, descriptive probability as described here may play an equally important role in statistical practice.

In most of the many academic disciplines that use observational data, the misuse and abuse of inferential statistics has been the norm for half a century or more. Thoughtful scholars have consistently criticized this misuse and abuse. David A. Freedman, one of the most thoughtful, compiled a compendium of such criticisms on pp. 212–217 of his posthumously published textbook [36]. But as the sociologist William S. Mason argued in 1991, in response to Freedman's powerful indictment of the inferential misuse of regression analysis, practitioners

[17] Huygens's second proposition, as translated from the Dutch by Hans Freudenthal [37]: "If I have an equal chance to get a or b or c, it is worth as much to me as though I had (a+b+c)/3." See also [63].

have a right to expect more than condescension from mathematical statistics [34, 50]. They need alternatives. Descriptive probability is one such alternative.

The frequentist vocabulary for statistical analysis is supported by an immense investment in research, training, and software. A comparable investment in the betting vocabulary will require decades of effort. It may, however, be the best contribution statisticians can make to the academic disciplines that use observational data. It may also be the best way to secure statistical insights within the emerging fusion of mathematical statistics with machine learning and other cultures in engineering and computer science that put more emphasis on prediction. The conundrums and phantasmagoria of frequentism, long tolerated within mathematical statistics, may be less sustainable in this wider culture.

# 6 Acknowledgments

This note was inspired by discussions in the Spring 2021 course on game-theoretic statistics that I taught jointly with Aaditya Ramdas and Ruodu Wang. I am grateful for their suggestions and for related discussions with Marshall Abrams, John Aldrich, Harry Crane, Nancy DiTomaso, Ruobin Gong, Zev Hirsch, Kitae Kum, Barry Loewer, Volodya Vovk, Sandy Zabell, and Snow Zhang.

# References

[1] John Aldrich. Burnside's engagement with the "modern theory of statistics". *Archive for History of Exact Sciences*, 63:51–79, 2009. 17

[2] Neal Alexander. What's more general than a whole population? *Emerging themes in epidemiology*, 12(1):1–5, 2015. 18

[3] David R. Bellhouse. *Leases for Lives: Life Contingent Contracts and the Emergence of Acturial Science in Eighteenth-Century England*. Cambridge, 2017. 15

[4] Richard A. Berk. *Regression Analysis: A Constructive Critique*. SAGE, 2004. 1, 18

[5] Richard A. Berk and David A. Freedman. Statistical assumptions as empirical commitments. In *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, pages 235–254. Aldine de Gruyter, Berlin, 2nd edition, 2003. 18

[6] Richard A. Berk, Bruce Western, and Robert E. Weiss. Statistical inference for apparent populations. *Sociological Methodology*, 25:421–458, 1995. 18

[7] Jacob Bernoulli. *Ars Conjectandi*. Thurnisius, Basel, 1713. 23

[8] Jacob Bernoulli. *The Art of Conjecturing, together with Letter to a Friend on Sets in Court Tennis*. Johns Hopkins University Press, Baltimore, 2006. Translation of [7] and commentary by Edith Sylla. 15

[9] Leo Breiman. Optimal gambling systems for favorable games. In Jerzy Neyman, editor, *Fourth Berkeley Symposium on Probability and Mathematical Statistics*, volume 1, pages 65–78. University of California Press, 1961. 6

[10] Leo Breiman. Statistical modeling: The two cultures (with discussion). *Statistical Science*, 16(3):199–231, 2001. 9

[11] Irwin Bross. Two-choice selection. *Journal of the American Statistical Association*, 45(252):530–540, 1950. 18

[12] Bernard Bru, Marie-France Bru, and Oliver Bienaymé. La statistique critiquée par le calcul des probabilités: Deux manuscrits inédits d'Irenée Jules Bienaymé. *Revue d'histoire des mathématiques*, 3:137–239, 1997. 3, 17

[13] Marie-France Bru and Bernard Bru. *Les jeux de l'infini et du hasard*. Presses universitaires de Franche-Comté, Besançon, France, 2018. 2 volumes. 14, 15, 17

[14] Santo F. Camilleri. Theory, probability, and induction in social research. *American Sociological Review*, 27(2):170–178, 1962. 17

[15] William G. Cochran. Relative accuracy of systematic and stratified random samples for a certain class of populations. *The Annals of Mathematical Statistics*, 17(2):164–177, 1946. 18

[16] Augustin Cournot. *Exposition de la théorie des chances et des probabilités*. Hachette, Paris, 1843. Reprinted in 1984 as Volume I (Bernard Bru, editor) of [17]. 16, 20

[17] Augustin Cournot. *Œuvres complètes*. Vrin, Paris, 1973–2010. The volumes are numbered I through XI, but VI and XI are double volumes. 23

[18] David R. Cox and David V. Hinkley. *Theoretical Statistics*. Chapman and Hall, 1974. 13

[19] Abraham De Moivre. *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play*. Pearson, London, 1718. Second edition 1738, third 1756. 15

[20] Morris H. DeGroot. A conversation with Charles Stein. *Statistical Science*, 1(4):454–462, 1986. 19

[21] Joseph L. Doob. Foundations of probability and its influence on the theory of statistics. In Donald B. Owen, editor, *On the History of Statistics and Probability*, pages 195–204. Dekker, New York, 1976. 20

[22] Anthony W. F. Edwards. *Likelihood: An Account of the Statistical Concept of* Likelihood *and its Application to Scientific Inference*. Cambridge, 1972. 13

[23] Stewart Ethier. *The Doctrine of Chances: Probabilistic Aspects of Gambling*. Springer, Berlin, 2010. 6

[24] Donald E. Farrar and Robert E. Apple. Some factors that affect the overhaul cost of ships: An exercise in statistical cost analysis. *Naval Research Logistics Quarterly*, 10(1):335–368, 1963. 18

[25] Hans Fischer. *A History of the Central Limit Theorem. From Classical to Modern Probability Theory*. Springer, 2010. 15

[26] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, 222(602):309–368, 1922. 7

[27] Ronald A. Fisher. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725, 1925. 5

[28] Ronald A. Fisher. Inverse probability and the use of Likelihood. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28(3):257–261, 1932. 18

[29] Ronald A. Fisher. Uncertain inference. *Proceedings of the American Academy of Arts and Sciences*, 71(4):245–258, 1936. 18

[30] Ronald A. Fisher. *Statistical Methods and Scientific Inference*. Hafner, 1956. Later editions in 1959 and 1973. 7

[31] Ronald A. Fisher. The nature of probability. *The Centennial Review of Arts & Science*, 2:261–274, 1958. 19

[32] Joseph Fourier. Mémoire sur les résultats moyens déduits d'un grand nombre d'observations. In Joseph Fourier, editor, *Recherches statistiques sur la ville de Paris et le département de la Seine*, pages ix–xxxi. Imprimerie royale, Paris, 1826. 2, 16

[33] Joseph Fourier. Second mémoire sur les résultats moyens et sur les erreurs des mesures. In Joseph Fourier, editor, *Recherches statistiques sur la ville de Paris et le département de la Seine*, pages ix–xlviii. Imprimerie royale, Paris, 1829. 2, 3

[34] David A. Freedman. Statistical models and shoe leather. *Sociological methodology (with discussion)*, 21:291–358, 1991. 18, 22

[35] David A. Freedman. Randomization does not justify logistic regression. *Statistical Science*, 23(2):237–249, 2008. 13

[36] David A. Freedman. *Statistical Models: Theory and Practice, Revised Edition*. Cambridge, 2009. 21

[37] Hans Freudenthal. Huygens' foundations of probability. *Historia Mathematica*, 7:113–117, 1980. 21

[38] Jules Gavarret. *Principes généraux de statistique médicale, ou développement des règles qui doivent présider à son emploi*. Bechet, Paris, 1840. 16

[39] Stephen M. Goldfeld and Richard E. Quandt. Nonlinear simultaneous equations: estimation and prediction. *International Economic Review*, 9(1):113–136, 1968. 18

[40] Irving J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3–4):237–264, 1953. 18

[41] Trygve Haavelmo. The probability approach to econometrics. *Econometrica*, 12(Supplement):1–115, 1944. 20

[42] Ian Hacking. *The Taming of Chance*. Cambridge University Press, 1990. 17

[43] Roger Hahn, editor. *Correspondence of Pierre Simon Laplace (1749–1827)*. Brepols, Turnhout, 2013. 2 volumes. 15

[44] David J. Hand. From evidence to understanding: a commentary on Fisher (1922) 'on the mathematical foundations of theoretical statistics'. *Philosophical Transactions of the Royal Society A*, 373:20140252, 2017. 5

[45] Edward Huth. Jules Gavarret's *Principes Généraux de Statistique Médicale*. *Journal of the Royal Society of Medicine*, 101(4):205–212, 2008. 16

[46] Marie-Françoise Jozeau. *Géodésie au XIXème Siècle: De l'hégémonie française à l'hégémonie allemande. Regards belges*. PhD thesis, Université Denis Diderot Paris VII, Paris, 1997. 8

[47] Lucien Le Cam. The central limit theorem around 1935 (with discussion). *Statistical Science*, 1(1):78–91, 1986. 17

[48] Erin Leahey. Alphas and asterisks: The development of statistical significance testing standards in sociology. *Social Forces*, 84(1):1–24, 2005. 17

[49] Paul Lévy. *Calcul de probabilités*. Gauthier-Villars, Paris, 1925. 20

[50] William M. Mason. Freedman is right as far as he goes, but there is more, and it's worse. Statisticians could help. *Sociological Methodology*, 21:337–351, 1991. 22

[51] J. Rosser Matthews. *Quantification and the Quest for Medical Certainty*. Princeton, 1995. 16

[52] Ernest Nagel. The meaning of probability (with discussion). *Journal of the American Statistical Association*, 31(193):10–30, 1936. 19

[53] John Neter. Applicability of statistical sampling techniques to the confirmation of accounts receivable. *The Accounting Review*, 31(1):82–94, 1956. 17

[54] University of Micigan Office of Diversity, Equity & Inclusion. Results of the 2016 University of Michigan faculty campus climate survey on diversity, equity & inclusion, 2017. `https://diversity.umich.edu/wp-content/uploads/2017/11/DEI-FACULTY-REPORT-FINAL.pdf`. 13

[55] Guy H. Orcutt, Harold W. Watts, and John B. Edwards. Data aggregation and information loss. *The American Economic Review*, 58(4):773–787, 1968. 18

[56] Marie-Vic Ozouf-Marignier. Administration, statistique, aménagement du territoire: l'itinéraire du Préfet Chabrol de Volvic (1773–1843). *Revue d'histoire moderne et contemporaine*, 44(1):19–39, 1997. 2

[57] Siméon-Denis Poisson. Observations relatives au nombre de naissances des deux sexes. *Annuaire le bureau des longitudes pour 1825*, pages 98–99, 1824. 16

[58] Siméon-Denis Poisson. Mémoire sur la proportion des naissances des filles et des garçons. *Mémoires de l'Académie royale des sciences*, IX:239–308, 1830. 16

[59] Richard M. Royall. Likelihood functions in finite population sampling theory. *Biometrika*, 63(3):605–614, 1976. 18

[60] Richard M. Royall. *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, 1997. 8, 11, 13

[61] Glenn Shafer. From Cournot's principle to market efficiency. In Jean-Philippe Touffut, editor, *Augustin Cournot: Modelling Economics*, pages 55–95. Edward Elgar, 2007. 20

[62] Glenn Shafer. On the nineteenth-century origins of significance testing and p-hacking, 2019. Working paper 55, `www.probabilityandfinance.com`. 8

[63] Glenn Shafer. Pascal's and Huygens's game-theoretic foundations for probability. *Sartoniana*, 32:117–145, 2019. 21

[64] Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication (with discussion). *Journal of the Royal Statistical Society: Series A*, 184(2):407–478, 2021. 2, 6, 14

[65] Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*. Wiley, New York, 2019. 2, 14

[66] William H. Starbuck. 60th Anniversary Essay: How journals could improve research practices in social science. *Administrative Science Quarterly*, 61(2):165–183, 2016. 1

[67] Stephen M. Stigler. Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1(3):359–378, 1986. 13, 15

[68] Stephen M. Stigler. *Statistics on the Table*. Harvard, 1999. Chapter 16, Daniel Bernoulli, Leonhard Euler, and maximum likelihood, pages 302–319. 13

[69] Marshall Stone. Mathematics and the future of science. *Bulletin of the American Mathematical Society*, 63:61–76, 1957. 20

[70] D. S. Tull, R. A. Boring, and M. H. Gonsior. A note on the relationship of price and imputed quality. *The Journal of Business*, 37(2):186–191, 1964. 17

[71] Richard von Mises. On the foundations of probability and statistics. *The Annals of Mathematical Statistics*, 12(2):191–205, 1941. 18

[72] Vladimir Vovk. Conformal testing in a binary model situation, 2021. Working paper 33, On-line Compression Modelling Project (New Series), `http://alrw.net`. 9

[73] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 2021. To appear, arXiv:1912.06116 [math.ST]. 6

[74] Abraham Wald. *On the Principles of Statistical Inference*. University of Notre Dame, 1942. 20

[75] William T. Ziemba. A response to Professor Paul A. Samuelson's objections to Kelly capital growth investing. *The Journal of Portfolio Management*, 42(1):153–167, 2015. 6