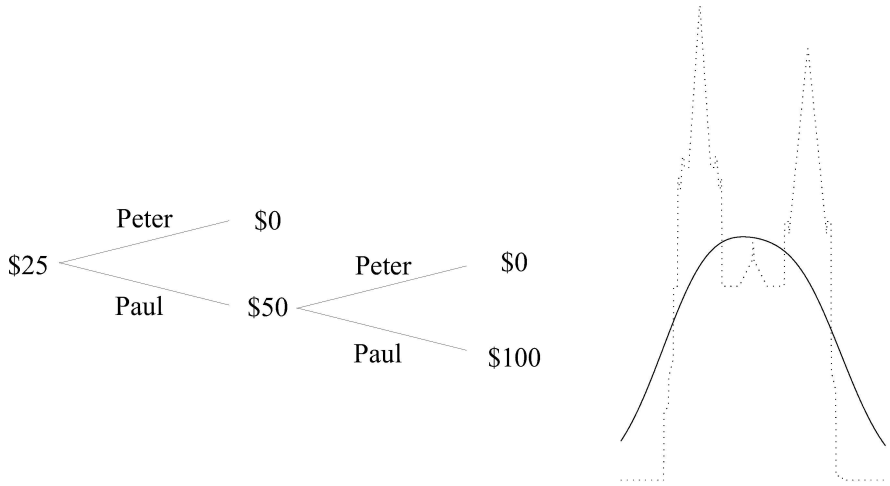
Bayesian, Fiducial, Frequentist

Glenn Shafer
Rutgers University
gshafer@rutgers.edu



The Game-Theoretic Probability and Finance Project

Working Paper #50

First posted April 30, 2017. Last revised December 31, 2017.

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

Abstract

This paper advances three historically rooted principles for the use of mathematical probability: the *fiducial principle*, *Poisson's principle*, and *Cournot's principle*. Taken together, they can help us understand the common ground shared by Bernoullians, Bayesians, and proponents of other types of probabilistic arguments. The paper also sketches developments in statistical theory that have led to a renewed interest in fiducial and Dempster-Shafer arguments.

1 Introduction

This paper is inspired by the recent emergence of a movement in theoretical statistics that seeks to understand and expand the common ground shared by Bernoullians and Bayesians and to reconcile their philosophies with R. A. Fisher's fiducial argument and its descendants, including the Dempster-Shafer theory of belief functions.[a]

I argue for three principles for the use of mathematical probability, principles that I believe will advance this search for common ground.

- The *fiducial principle*: All use of mathematical probability is fiducial. It requires, that it is to say, a decision to trust particular probabilities in a particular situation, even though these probabilities are initially purely hypothetical, theoretical, subjective, or derived from other situations, which can never be identical to the situation at hand in all respects.

- *Poisson's principle*: Even varying probabilities allow probabilistic prediction. The law of large numbers, for example, does not require independent identically distributed trials. Moreover, the predictions may concern averages or other statistics rather than the frequencies of particular events. An interpretation of probability that emphasizes its predictions should not, therefore, be called *frequentist*.

- *Cournot's principle*: Probability acquires objective content only by its predictions. To predict using probability, you single out an event that has high probability and predict it will happen. Or, equivalently, you single out an event that has small probability and predict it will not happen.

Each of these principles has venerable historical roots. Each is, in some sense, a truism. But the three principles are generally left in the background in philosophical discussions of statistical testing, estimation, and prediction. By making them explicit and salient, we can dispel some of the misunderstandings that have kept Bernoullian and Bayesian statisticians and other philosophers of probability talking past each other.

The fiducial principle identifies a feature common to fiducial and Bernoullian statistical practice that even Bayesians cannot completely escape. Fisher's fiducial argument singles out particular probability statements about the relation between two unknown quantities before either is known and continues to consider them valid after one of the quantities is observed. Bernoullians find a different way to have continued confidence in such probability statements after one of the quantities is observed. Bayesians recoil at this, but they too single out particular probability statements to continue to trust, and they too can never take account of all the evidence they observe.

[a] Non-Bayesian methods of statistical estimation and testing are now often called *frequentist*. Following Francis Edgeworth, Richard von Mises, Arthur Dempster, and Ian Hacking, I am calling them instead *Bernoullian*, in honor of Jacob Bernoulli.

Poisson's principle helps us see past the widely held but fallacious thesis that Bernoullian statistical theory equates probability with frequency. Stochastic processes, used as models by both Bernoullians and Bayesians, can be far removed from the picture of repeated trials under constant conditions, and an individual probability in a stochastic process may have nothing to do with the frequency of any event.

Cournot's principle is integral to Bernoullian statistics, where it is used for prediction (we single out events with high probability and predict that they will happen) and testing (the model or theory is discredited when such a prediction fails). But many statisticians who call themselves Bayesian also rely on this logic of testing.

In the next section I review the historical context and recent developments. In subsequent sections (Sections 3, 4, and 5), I discuss the three principles in turn. I conclude with a brief summary (Section 6). Throughout, I try to keep the mathematical exposition as elementary as possible

There are two appendices. Appendix I reviews the history of the adjectives *Bayesian*, *Bernoullian*, and *fiducial* and related terms. Appendix II briefly discusses the game-theoretic understanding of mathematical probability that informs some of the ideas presented here.

There are two groups of notes. Footnotes, identified with Latin letters, provide information that may help some readers follow the text. Endnotes, numbered with Arabic numerals, provide supplementary information, mostly historical, and appear at the end of the paper, before the references.

2 Context

In this section, I review Fisher's framework for theoretical statistics, his fiducial argument, its principal difficulties, the Dempster-Shafer generalization, and the crisis of Bayesian practice that has led to renewed interest in fiducial arguments.

2.1 Fisher's framework

When R. A. Fisher's work began to attract widespread attention in the 1920s, the British biometric school, led by Karl Pearson and collaborators such as William S. Gosset and George Udny Yule, had already established international leadership in mathematical statistics. Their contributions included new models and methods of estimation and testing, as well as the introduction of correlation and regression and new methods for analyzing time series. Fisher's further contributions included distribution theory for numerous small-sample statistics, the theory of maximum likelihood, and methods for designing experiments and analyzing variance. One of Fisher's most influential contributions was his 1922 article "On the mathematical foundations of theoretical statistics" [90]. This article is most often remembered for its theory of maximum likelihood and the concepts of consistency, efficiency, and sufficiency, but its most deeply influential contribution may have been its doctrine that theoretical statistics begins with

a parametric statistical model, say an indexed class of probability distributions $\{P_\theta\}_{\theta\in\Theta}$, and that the task of theoretical statistics is to use a random sample to estimate the *parameter* θ.[b]

Fisher explained this abstract starting point by saying that theoretical statistics begins after the practical statistician has specified "a hypothetical infinite population, of which the actual data are regarded as constituting a random sample." The theoretician's task is to estimate from this data the "law of distribution of this hypothetical population", which "is specified by relatively few parameters". He assumed that the number of observations in the random sample was large relative to the number of parameters (the dimensionality of θ), so that this task can be thought of as "data reduction".[1]

Fisher's framework is now so taken for granted, and seems so helpful for understanding aspects of statistical theory in the century before Fisher as well as the century after, that it is difficult to analyze its originality. It reflects much of what came before, from Jacob Bernoulli's estimation of the probability of an event to Karl Pearson's fitting of frequency curves, but it abstracted from previous practice in at least three original ways:

1. Most novel, perhaps, was the unqualified assumption that the class of probability distributions $\{P_\theta\}_{\theta\in\Theta}$ is always known – that it has been put in place by the "practical statistician". The theory of errors, as formulated by Laplace and Gauss, had not assumed that we know the probability law for the errors. Laplace's celebrated normal approximation of 1810, now seen as an early version of the central limit theorem, was important to Laplace and his 19th-century successors precisely because it allows us to draw inferences when these probabilities are not known, provided we have many independent measurements.[2]

2. The level of abstraction was new, at least for most of Fisher's readers.[3] In the 19th century, the theory of statistical estimation was concerned with estimating quantities that had a concrete meaning in the world independent of any probabilities. Writing in 1884 [75], Francis Edgeworth distinguished between real quantities, such as a star's position, and fictional quantities, such as the average flowering-time of a plant. But even Edgeworth's fictional quantities were described directly in terms of features of the world. Karl Pearson's work with W. F. R. Weldon shifted the attention of statisticians from estimating such self-standing quantities to fitting frequency distributions – the distribution of flowering-time for a plant or the distribution of forehead-size for a species of crab [213]. Pearson addressed this problem by inventing classes of distributions with adjustable constants and estimating the constants from Weldon's data.

[b] Here θ may be a single real number or a vector of real numbers; thus Θ is a subset of the real numbers or a subset of a Euclidean space. When θ is a vector, it is now customary to call both it and its components parameters. A real-valued function $h(\theta)$, such as $\theta_1 - \theta_2$ when $\theta = (\theta_1, \theta_2)$, may also be called a parameter, but for clarity I will instead call it a *feature* of the parameter.

Fisher made this picture abstract, calling Pearson's frequency curves laws of probability and calling Pearson's frequency-constants *parameters*.[4]

3. The framework was also original and powerful by virtue of its narrowness – what it left out. It put the random sample (independently and identically distributed observations) at the center of theoretical statistics, relegating to a peripheral role most of the statistical theory of the preceding century, including time series and least squares, not to mention topics to which Fisher himself was to make pathbreaking contributions: significance testing, multiple regression, randomization, and the design of experiments. The narrowness can be understood in the context of Fisher's leadership struggle with Pearson, for Pearson and his fellow biometricians were emphasizing random sampling from biological populations. But most statistical work at the beginning of the 20th century was in fields such as economics, demography, insurance, and meteorology, where time series are central. Even Pearson, Gosset, and Yule contributed to the theory of time series.[5]

For many older statisticians, Fisher's pronouncements concerning the task of theoretical statistics sounded ridiculous.[6] But time series as a branch of probability theory, the field of study we now call *stochastic processes*, was in its infancy in 1922.[7] Fisher's narrowing of the scope of theoretical statistics to the random sample enabled him and his immediate successors to provide a firmer foundation for the subject using the existing probability calculus. The success of this mathematical work has kept the random sample at the center of mathematical statistics even to this day, sometimes in ways we may not recognize. Today we are accustomed to parametric statistical models in which the P_θ are probability laws for a stochastic process for which there will be only a single observation, or in which the number of parameters far exceeds the number of observations, so that an estimate of θ is hardly a data reduction. Yet there is still a temptation to suppose that P_θ's probabilities must be understood in terms of hypothetical repeated draws from a hypothetical population. (See the discussion of the "repeated sampling principle" on page 36 below.)

Fisher's abstract framework also subtly changed the relationship between direct (Bernoullian) and inverse (Bayesian) probability. After the work of Laplace and Gauss in the early 19th century, the two methods had co-existed for a century, often peacefully. Inverse probability was attractive to many mathematicians, but because probabilities for observations given by causes (the P_θ in Fisher's formulation) were usually considered unknown, and Gauss's direct probability argument (now called the *Gauss-Markov theorem*) applied even to relatively small samples, direct arguments were seen more often in statistical practice.

Laplace had introduced the distinction between direct and inverse probability in 1774 ([131], Section II). He explained that there are two classes of problems in probability theory: direct problems, in which we seek the probabilities of events from causes, and inverse problems, in which we seek the probabilities of causes from events.[8] This distinction between *cause* and *event* did the same

work as Fisher's distinction between parameter and data. A cause is a possible value of Fisher's θ (an element of Θ). An event is the data, say $x_1, \ldots, x_n$.

Laplace advanced a simple principle for solving inverse problems: in light of an event, the probability of each possible cause should be proportional to the probability the cause gives the event.[9] Let us write $P_\theta(x_1, \ldots, x_n)$ for P_θ's probability for $x_1, \ldots, x_n$ (or for its density if the possible values for the data vary continuously). If Θ is finite or countable, then Laplace's principle says that

$$\text{posterior probability}(\theta) := \frac{P_\theta(x_1, \ldots, x_n)}{\sum_{\tau \in \Theta} P_\tau(x_1, \ldots, x_n)},$$

where *posterior* means posterior to seeing the data $x_1, \ldots, x_n$. If the parameter varies continuously, then the principle says that

$$\text{posterior density}(\theta) := \frac{P_\theta(x_1, \ldots, x_n)}{\int_{\tau \in \Theta} P_\tau(x_1, \ldots, x_n) d\tau}.$$

Fisher called $P_\theta(x_1, \ldots, x_n)$, considered as a function of θ with $x_1, \ldots, x_n$ fixed, the *likelihood function*. So in Fisher's language, Laplace's principle tells us to obtain probabilities by normalizing the likelihood function.[c]

For more than thirty years, Laplace was unable to evaluate the integrals involved in applying his inverse principle to errors of observations. But his 1810 approximation result enabled him to do so in the case of many observations, and he obtained not only inverse-probability solutions but also direct-probability solutions. They were more or less identical, and they justified the method of least squares. They assumed that the error distribution was unknown but was symmetric around zero and essentially bounded.

The nature of the near identity between inverse and direction solutions is more readily explained for the binomial problem considered by Bernoulli and Bayes than for the errors-in-observation problem that Laplace finally conquered in the 1810s. In the binomial problem, the inverse and direct solutions both say that

$$P\left(\left|\frac{y}{n} - p\right| \le 2\sqrt{\frac{\frac{y}{n}(1 - \frac{y}{n})}{n}}\right) \approx 0.95, \tag{1}$$

where p is the unknown probability of an event, n is large and y is the number of times the event happens in n trials. In the direct-probability interpretation, p is fixed, and 0.95 is the probability that y satisfies the inequality. In the inverse-probability interpretation, y is fixed and 0.95 is the probability that p satisfies the inequality. These results were already more or less available in the 1770s; the direct-probability result, which follows from the central limit for the binomial established by De Moivre in 1733, was spelled out by Lagrange in 1776

[c] Laplace's principle has also been called Bayes's rule, but Bayes's formulation of it, published in 1763 [7], was not known to Laplace in 1774, and Bayes's argument for the rule was not influential until the second half of the 20th century. In 1922, the principle was still known in English as the principle of inverse probability.[10]

(see for example [114]). The inverse-probability result can be obtained from the methods in Laplace's 1774 article (see [108]).

Once Laplace had obtained his direct-probability justification for least squares, he and others using the theory of errors preferred it over his inverse-probability justification. Anders Hald reports that Laplace never used inverse probability after 1811, and Gauss never used it after 1816. The direct argument was sufficient and, as Gauss once wrote, less metaphysical. In 1823, moreover, Gauss gave a simpler direct-probability argument, called the *Gauss-Markov theorem* in later textbooks, that accomplishes nearly as much even when the number of observations is not so large.[11]

By the middle of the 19th century inverse probability was the object of much explicit criticism, beginning most notably in Antoine Augustin Cournot's *Exposition de la théorie des chances et des probabilités* in 1843 [32]. Cournot defended direct-probability arguments, but others challenged the validity and usefulness of the entire probability calculus, and most mathematicians continued to consider the inverse principle an integral part of that calculus. Throughout the second half of the 19th century and into the 20th, most treatises on probability included the inverse principle, even as more applied work on the theory of errors and mathematical statistics generally relied on probable errors[d] interpreted in terms of direct probability.[12]

As he stated it in 1774, Laplace's inverse principle made no mention of prior probabilities. It assumed, implicitly, that prior probabilities are equal or uniformly distributed. Some of Laplace's 1774 examples show that he understood that unequal prior probabilities are sometimes needed, but he evidently thought it unnecessary to mention them in the statement of the principle, as they could be introduced in other ways. Only in 1814, in his *Essai philosophique sur les probabilités* [133], did he did finally mention prior probabilities in his statement of the principle.[13]

In the final decades of the 19th century, unequal prior probabilities became more prominent in discussions of inverse probability. In 1884, for example, Frances Edgeworth devoted an article to them [76]. But even at the beginning of the 20th century, the terms *inverse probability* and *Bayes's rule* were often taken to refer to a formula in which prior probabilities do not appear and hence a uniform distribution of probabilities is assumed.

Karl Pearson worked in the 19th-century tradition of co-existence between direct and inverse probability, calculating standard deviations for his estimators using whatever method, direct or inverse, seemed most convenient. He also relied on inverse probability in his philosophy of science. In *The Grammar of Science*, the influential book on the philosophy of science he first published in 1892 [172], he advised readers who wanted to learn about probability to consult Thomas Galloway's 1839 treatise [100], which taught inverse probability

[d] The notion of a *probable error* (*error probabilis* in Latin) was introduced by Bessel in 1818 [14]. The probable error of an estimate $\hat{\theta}$ of a quantity θ is the number m such that $\hat{\theta} - m \leq \theta \leq \hat{\theta} + m$ with probability 50%. The standard error σ of the estimate, introduced by Karl Pearson, is now more familiar; $m \approx 0.67\sigma$ when the estimate is normally distributed.

as developed by Laplace and Poisson. He quoted Francis Edgeworth [76] in defense of uniform prior distributions: "the assumption that any probability-constant about which we know nothing in particular is as likely to have one value as another, is grounded upon the rough but solid experience that such constants do as a matter of fact as often have one value as another."[14]

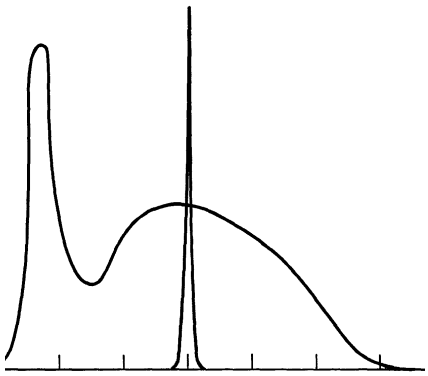
Fisher upended this co-existence of direct and inverse probability. His theory of estimation (sufficiency, maximum likelihood, etc.) used direct probability, and he forcibly criticized inverse probability on the grounds that its results depend on the scale for the parameter (or, equivalently, on the choice of prior probabilities). But his very framework, especially his declaration that theoretic statistics is concerned only with estimation using random samples (not with testing, for example) and his assumption that the probabilities given the parameter values are known, made inverse probability appear as a completely general and simple way of doing theoretical statistics: just normalize the likelihood function, first multiplying by a factor representing prior probabilities if needed. Little wonder that Fisher denounced this alternative so fiercely.[15]

2.2 Bayesianism

From the 1920s into the 1960s, the development of mathematical statistics was led by Fisher and then by Jerzy Neyman, Egon S. Pearson, and Abraham Wald. Neyman and Egon S. Pearson were initially more interested in inverse probability than Fisher, but they and Wald eventually agreed with Fisher that Bayes's theorem was of little use in statistical analysis because of its reliance on prior probabilities. Neyman, Pearson, and Wald departed from Fisher, however, by emphasizing decisions based on statistical evidence. This emphasis, bolstered by the development of game theory and decision theory in the late 1940s and early 1950s, led to renewed acceptance of subjective probability and subjective expected utility. This led in turn to the development, beginning in the late 1950s, of a new school of thought that called itself Bayesian. Influenced by decision-theoretic arguments that suggested the need for subjective probabilities, and appealing to earlier work on subjective probability by Frank P. Ramsey and Bruno de Finetti, most of the new Bayesians considered it unnecessary to justify uniform probabilities as an expression of ignorance or rough past experience. Each person should settle on their own subjective probabilities.[16]

The vast majority of the new Bayesians were never as thoroughly subjective, however, as their rhetoric suggested. While insisting on the subjective nature of the prior probabilities for θ, they continued to interpret the probability distributions P_θ objectively, just as Fisher and Neyman had done. Like Bernoulli, Laplace, Gauss, and almost all mathematical statisticians since, they thought unknown probability laws were features of the world. To see why this is so, consider the interpretation that Bruno de Finetti, known for his uncompromising subjectivism since the 1930s, proposed for the Fisherian framework in 1953 [52]. The probabilities given by P_θ, he proposed, are subjective probabilities everyone would have if they knew the value of θ. Such conditional opinions might indeed be understood in a purely subjective way when θ has some reference in

Figure 1: Example used by Edwards, Lindman, and Savage (1963) to illustrate the principle of stable estimation.



The bimodal curve is the prior density. The spiked curve is the likelihood function, which can be expected, when there many observations, to have approximately the shape of a normal density. A normal density is effectively zero outside an interval extending a few standard deviations from its peak. So the posterior density, which is proportional to the product of the prior density and the likelihood, will also be zero outside this interval. When their are many observations, the interval is very narrow. On the reasonable assumption that the prior density is approximately constant over this narrow interval, it will make little difference.

---

the world aside from the probabilistic predictions it makes, as when it represents the fraction of balls in an urn or the true value of a quantity being measured. But in general, there is no such reference in Fisher's picture. Instead, θ is merely a constant that we adjust to fit the data. In this case, statisticians must put a Bernoullian gloss on de Finetti's formulation: θ is the hypothesis that P_θ gives accurate frequencies or withstands gambling strategies. You would adopt P_θ as your subjective probability distribution if you knew this hypothesis were true, but the hypothesis itself is an objective interpretation of P_θ's probabilities.[17]

Far more often than not, Bayesian statisticians also hope that the conclusions of their analyses will be at least approximately valid from the Bernoullian point of view. Typically they have a plan for selecting and announcing a set of values of θ that has posterior probability near one, and they want each P_θ to predict that this plan will be successful. In other words, for each θ, they want P_θ to give a probability close to one that θ will be in the announced set. They often express this by saying that they want their Bayesian analyses to have good "frequency properties".

When sample sizes are large enough, the prior probabilities are smooth, and

θ is one-dimensional or perhaps two-dimensional, Laplace's approximation of 1810 can be deployed to show that the prior probabilities do not matter very much, and that Bayesian analyses will have Bernoullian properties. This was well understood in the 19th century, and it was textbook fare by the early 20th century.[18] Then, as now, it was expressed in various ways.

1. We can say simply that the prior probabilities do not matter much, and hence that we can simplify by using uniform prior probabilities, as in Laplace's and Bayes's original formulations. In the early 1960s, Leonard J. Savage made this point, calling it the *principle of precise estimation* [184] or the *principle of stable estimation* [82]. The latter article, which Savage published in 1963 with co-authors Ward Edwards and Harold Lindman, used Figure 1 to illustrate the principle of stable estimation.

2. We can explain in more detail that the posterior probability for θ will be approximately normal, with mean and variance that can be estimated from the data and do not depend on the prior probabilities. This was emphasized, for example, by Harold Jeffreys in 1939 [122].

3. A Bayesian can explain the acceptability of Bernoullian analyses by showing that they approximate the Bayesian analysis with a smooth prior. John W. Pratt made this point in 1965 [177].

4. A Bernoullian can explain the acceptability of Bayesian analyses by showing that they approximate Bernoullian analyses. This view was taken in 1843 by Cournot ([32], Section 95).

5. Ignoring direct Bernoullian analyses, Bayesians can claim Bernoullian or "frequentist" properties for their results.

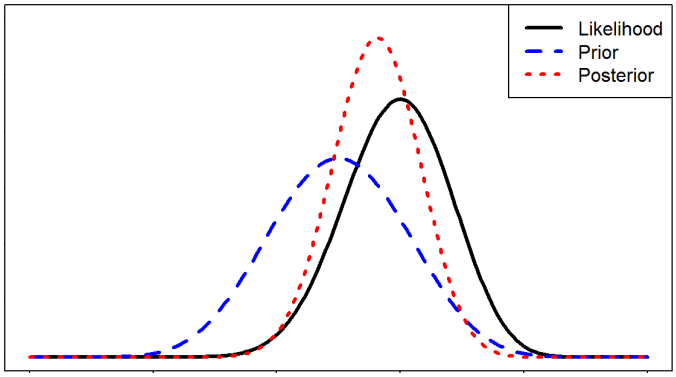
In the first flush of their new faith in the 1960s, the new Bayesians emphasized Points 1 and 3. But in more recent decades, even as the ranks of mathematical statisticians calling themselves Bayesians has swelled, their emphasis has shifted from defending subjective probability to seeking Bayesian procedures with Bernoullian properties.[19]

When the sample size is not large enough for the principle of stable estimation to be applicable, Bernoullian properties may be elusive, but Bayesians can emphasize that the posterior distribution for θ is always a legitimate compromise between the prior distribution and the likelihood. This compromise is often illustrated with pictures like Figure 2, where the prior density, the likelihood, and the posterior density are all unimodal.

2.3 Fiducial and Dempster-Shafer arguments

In 1922, Fisher's rejection of inverse probability led him to conclude that a random sample from a parametric model does not justify stating probabilities about θ. You have only a likelihood function for θ, and this is not a probability distribution. It tells you the relative likelihood of two different values of θ (by

Figure 2: The compromise between prior and likelihood.



The posterior density is a compromise between the prior density and the likelihood, and so its mode falls between their modes. Examples of this type appear in many expositions of Bayesian statistics.

Fisher's new definition of likelihood!), but its values do not sum or integrate to one, and you cannot sum or integrate over a subset of Θ to get a probability for that subset.

The proposition that one cannot make probability statements about a parameter based on observations was contrary to statistical tradition, however. Statisticians had long used both direct and inverse large-sample methods to make such statements. Fisher's admirers were soon using direct probability to calculate small-sample error limits for parameters, in the same spirit as Laplace's large-sample error limits [2]. In 1930, Fisher convinced himself that this is sometimes legitimate. He called the probabilities thus obtained for θ *fiducial*.

Suppose, to consider the simplest example, that we are about to measure a quantity θ. Our measuring apparatus makes errors. If we write x for the result of the measurement and u for its error, then

$$x = \theta + u. \qquad \text{(structural equation)}$$

We can also write this as

$$\theta = x - u \qquad \text{(fiducial equation)}$$

or

$$u = x - \theta. \qquad \text{(pivot equation)}$$

Suppose we have a probability distribution P for the error u, based perhaps on past experience, and suppose P is the normal distribution with mean zero and variance one. If we observe $x = 2.3$, say, then our fiducial equation is

$\theta = 2.3 - u$. If we continue to trust the probability distribution P for u (this thought animates all Bernoullian arguments), then by the rules of probability, θ is normal with mean 2.3 and variance one. This is θ's *fiducial probability distribution*.

This example is relatively appealing, because it supplies a story about the origin of the Fisherian model (x is normal with mean θ and variance one) and the fiducial equation ($\theta = x - u$), a story that harks back to Laplace's and Gauss's error theory. In general, however, we begin with only with a Fisherian model, supposedly supplied by the practical statistician. We must then invent a fiducial equation that fits this model, perhaps choosing it somewhat arbitrarily from multiple possibilities.

Fisher developed his fiducial argument by giving examples, not by laying out a general theory. Usually he looked for a pivot equation, an equation of the form

$$u = \psi(x, \theta) \tag{2}$$

such that (1) u's probability distribution is the same for all θ, (2) ψ depends on the data x only through the minimal sufficient statistic,[e] and (3) the equation can be solved uniquely for θ so as to obtain a fiducial equation

$$\theta = H_x(u). \tag{3}$$

Other authors have often preferred to begin with a structural equation

$$x = G(\theta, u) \tag{4}$$

that can be solved uniquely to obtain a fiducial equation. This starting point is attractive because it allows us to imagine that it encodes an origin story for the parametric model, as in our example of a single measurement with error.[20]

The quantity u in Equation (2), the pivot equation, is called the *pivot*. In order to obtain direct-probability statements about θ by inverting (2), it is sufficient that the pivot's probability distribution be the same, at least approximately, for all θ. As we have noted, Lagrange obtained such a direct-probability statement for the large-sample binomial problem in 1776. Cournot explained the logic of the inversion in 1843 (see Section 3.1). This logic does not require that u depend only on the minimal sufficient statistic, and it does not even require that the pivot equation be uniquely solvable for θ, but the direct-probability statements do not constitute probabilities for θ in the usual sense. In 1937, Neyman underlined this point by calling them degrees of *confidence* [163]. Fisher, on the other hand, believed that in his examples, where u depends only on a sufficient statistic and the dependence is continuous and strictly monotonic, a probability obtained from Equation (3) is a probability like any other probability.

[e] Fisher called any function of the observations $x_1, \ldots, x_n$ a *statistic*. He called a statistic *sufficient* if its probability distribution under P_θ is the same for all θ. He believed that a sufficient statistic captures all the information about θ that is provided by the observations. Thus a minimal sufficient statistic represents the maximal reduction of the data that retains all its information about θ.

Fisher's view was at first difficult to refute, because he did not lay out what he expected from a probability.[21] But puzzles began to emerge as soon as Fisher and his followers began to consider examples in which θ is multidimensional. Particular attention became focused on the problem of two normal distributions with unknown means μ_1 and μ_2 and unknown variances. Given independent observations from the two distributions, what should we say about the difference $\mu_1 - \mu_2$ (this is the *Behrens-Fisher problem*) or about the ratio μ_1/μ_2 (this is the *Fieller-Creasy problem*) [236]? It was not always clear how to answer these questions on Fisher's principles, and proposed answers had properties most statisticians were loath to accept. You could use the data to make money betting against some of these answers.

Fisher's contributions had earned him so much prestige and so many loyal followers that he was able to deflect and ignore criticisms of his fiducial ideas in the 1930s and 1940s [246], but he paid more attention in the mid-1950s, when Georges Darmois, who had championed his ideas in France, showed him a critique by the Russian mathematician Andrei Kolmogorov in 1942 [129].[22] Kolmogorov noted that Fisher's fiducial inversion can produce "probabilities" that do not have a property that Richard von Mises had called the *irregularity axiom*.[23] This axiom says that the information you have when you make bets at the odds set by a declared probability should not enable you to pick out trials on which the long-run frequency is different from that probability. Apparently after reading Kolmogorov's critique (in a French translation supplied by Darmois), Fisher gave this property his own name; he called it the *absence of recognizable subsets*. But he did not abandon his intuition about sufficiency. In his 1956 book, *Statistical Methods and Scientific Inference* [94], he claimed, without proof, that his fiducial probabilities, when based on sufficient statistics (perhaps conditioned when appropriate on statistics whose distribution is the same for all values of the parameter) did not admit recognizable subsets. He was flatly wrong. In 1963, a year after Fisher's death, Robert J. Buehler and A. P. Feddersen closed the book on Fisher's argument by showing that even intervals based on Student's t-distribution, the most basic example of Fisher's theory, admitted recognizable subsets [29, 188, 246].

Fisher had advanced his argument only for models with continuous observations (otherwise the structural equation cannot be inverted to obtain a fiducial equation), but in 1957 he suggested that something similar could be done with discrete models such as the binomial, even if this did not produce precise probabilities.[24] Arthur Dempster took up this idea in a series of articles in the 1960s, giving methods for obtaining upper and lower probabilities for both continuous and discontinuous models [58, 59, 60, 61, 63, 65, 66]. Dempster used a structural equation of the form (4), but he did not require that it be uniquely solvable for θ when x is fixed so as to yield a fiducial equation of the form (3). He accommodated values of u for which there are multiple solutions by mapping u to the set of solutions (thus describing a random subset of Θ rather than a random point in Θ), and he eliminated u for which there are no solutions by conditioning on those for which there are, in the manner of Bayes.

Dempster's argument produces the same results as Fisher's fiducial argu-

ment in some problems, such as the simple measurement problem discussed on page 10, and it is often considered a generalization of that argument. It has a different starting point, however. Fisher started with a parametric model and added structure to it by defining a pivot in terms of the model. Dempster began instead with a parameter space Θ and a probability distribution for a variable u unrelated to Θ. He then specified a structural equation $x = G(\theta, u)$ and assumed that the observations were obtained from θ and u via this equation. A parametric model can be obtained from this picture by fixing θ, but it is not basic. Because Dempster's interpretation of the probability distribution u was subjectivist, he did not ask for his inferences to have Bernoullian properties with respect to this parametric model.

While being related to the fiducial argument, Dempster's method also constituted a generalization of the Bayesian calculus, and like the Bayesian calculus it can be used outside the Fisherian framework. I presented it in this general way in my 1976 book, *A Mathematical Theory of Evidence* [190]. In the 1980s it was widely used in artificial intelligence under the name *Dempster-Shafer theory* [245].

Dempster-Shafer belief functions have found their greatest use in domains where statistical models have little traction because it is impossible, impractical, or implausible to model in advance the evidence we might obtain, but where we nevertheless want to quantify and formally combine various items of evidence, including evidence that provides little or no support for either side of some questions being considered. This includes domains such as financial auditing, assurance services, the assessment of intelligence, and judicial deliberation.[25] The most important tools in these domains are the rule of combination, introduced by Dempster in his articles in the 1960s, and belief-function discounting, introduced in my 1976 book.[f] Because there is no parametric model in these applications, the issue of Bernoullian properties with respect to a parametric model does not arise.

In 1982 [194], I argued that a Fisherian model and accompanying observations may not provide enough information to permit an analysis using belief functions; what is missing is the evidence that justifies the model. In cases where we can say something about this missing evidence (as when we have a story justifying a particular probability distribution for an anticipated error), it may be possible to model it in ways more amenable to persuasive belief-function analysis. Dempster has repeatedly made related arguments, beginning in the foreword that he wrote for my 1976 book. In a recent article [70], he has pointed out that once Bayesian models and analyses are re-expressed in Dempster-Shafer terms (and thus given the additional structure represented by a structural equation), it becomes clear that both the prior distribution and the likelihood function can be weakened to reflect the weakness or absence of underlying evidence.

[f] Both ideas had already been used used already by Jacob Bernoulli and George Hooper in the 17th century [191, 196], and I learned from their work as well as from Dempster.

2.4 The 21st century Bayesian crisis

By the 1980s, subjective Bayesianism was gaining ground in applied statistics because of the increasing size and complexity of datasets and the concomitant complexity of the Fisherian parametric models proposed for their analysis. Jerzy Neyman and E. S. Pearson's ideas for choosing among Bernoullian tests and confidence intervals, so popular since the 1930s for problems in which θ was a single quantity or a low-dimensional vector, proved less helpful for these more complex models. Since the 1980s, increasing computational power and more sophisticated computational methods, especially Markov chain Monte Carlo, have made Bayesian analyses more and more practical. Today Bayesian methods are widely used, and their position vis-à-vis Bernoullian methods is at least as strong as it was in the 19th century.

As data and models have become even more complex, however, Bayesian analyses have become less transparent, and the Bayesian procedure has lost the simple properties that made it so attractive and persuasive in the last decades of the 20th century. We are often interested in a particular feature $h(\theta)$ of the complex multi-dimensional parameter θ, or in a few such features, and in many cases the Bayesian procedure for obtaining probabilities for particular features may not have desired Bernoullian properties or even produce a reasonable compromise between the prior and the likelihood.

As statisticians have long understood, the likelihood function for a multi-dimensional parameter, even when there are so many observations that it is concentrated on a relatively small region of the parameter space Θ, may not be concentrated with respect to a particular feature $h(\theta)$ of interest. In relatively simple models this problem can be detected, and the model can be modified to remedy it.[g] But in complex models the problem may be difficult to detect.

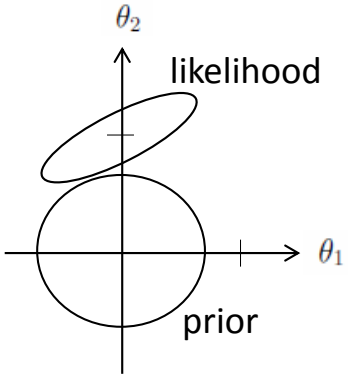
Moreover, the geometry of high-dimensional spaces makes Edgeworth's and Pearson's notion of diffuse and unopinionated prior probabilities grounded on "rough but solid experience" elusive when $\theta = (\theta_1, \ldots, \theta_k)$ and k is large. A diffuse prior distribution for $(\theta_1, \ldots, \theta_k)$ expresses strong opinions about some features. A uniform distribution on the cube $[-K, K]^k$, where K is large, for example, suggests that $h(\theta) := \sum_{i=1}^{k} \theta_i^2$ is very large. Yet a prior distribution concentrated around an anchor point expresses equally strong opinions; the ball $\{\theta \in \mathbb{R}^k | \sum_{i=1}^{k} \theta_i^2 \leq 1\}$ is a very small part of the cube $[-1, 1]^k$.

It follows that if we avoid extremely diffuse priors because of their extreme opinions, we cannot provide a prior density that will be relatively uniform, no matter how the likelihood function comes out, over the multi-dimensional region of values where this likelihood has non-negligible values. Thus the principle of stable estimation, as illustrated in Figure 1, does not generalize beyond the case where Θ has just a few dimensions.

It is even very possible that the prior distribution will give little probability

[g] One very widely understood example is that of "multicollinearity" in multiple regression, where the estimation of two regression coefficients is highly uncertain because of the independent variables associated with them are almost linearly related. One remedy is simply to omit one of these variables.

Figure 3: How a Bayesian posterior can fail to be a compromise between the prior and the likelihood. (Example suggested by Min-ge Xie.)



The circle is a contour for the prior density. The tilted ellipse is a contour for the likelihood function. Both suggest that 0 is the most likely value for θ_1.

The posterior density, being proportional to product of the prior and the likelihood, is greatest in the region where the two contours come closest, suggesting a negative value for θ_1.

Far from being exceptional, this failure to compromise arises for some feature $h(\theta_1, \theta_2)$ whenever the prior density and likelihood function are tilted with respect to each other.

to the region where the likelihood is concentrated, and in this case we cannot even count on the posterior being a compromise between the prior and the likelihood for particular features that interest us. Even in two dimensions, as Min-ge Xie has pointed out to me, the likelihood function and the prior density will typically be tilted with respect to each other, and then there will be real-valued functions $h(\theta)$ for which the posterior density, instead of being a compromise between the prior and the likelihood, falls to the same side of both of them.

Figure 3 illustrates this point. Here we have a two-dimensional parameter $\theta = (\theta_1, \theta_2)$. The prior density is centered on $\theta_1 = \theta_2 = 0$, but the maximum-likelihood estimate is $\theta_1 = 0$, $\theta_2 = 1$. The posterior density is greatest in the region where high contours of the prior density are closest to high contours of the likelihood function. In the case of θ_2, this results in a compromise between the prior and the likelihood – a posterior mode between the prior mode 0 and the maximum-likelihood estimate 1. But because the likelihood function is tilted, we do not obtain a compromise for θ_1; the prior mode and the maximum-likelihood estimate are both zero, but the posterior mode is negative.

The picture in Figure 3 can arise in many ways. Suppose, for example, that θ_1 and θ_2 are independent and standard normal under the prior, and suppose the likelihood arises from a single bivariate normal observation (x, y), x having mean θ_1 and variance 1, y having mean θ_2 and variance 0.2, and the two having correlation 0.8. A standard calculation shows that the posterior is bivariate normal, with mean -0.08 for θ_1 and mean 0.97 for θ_2. In this case, the tilt in the likelihood arises because of the prior assumption that the correlation 0.8, and one might object to giving this particular value for the correlation probability one. But if there were many observations instead of a single observation, we would expect a tilt just from the randomness of the data. In general,both

because of the randomness of the data and the complexity of the model, we can expect that the contours of the likelihood will be tilted with respect to the contours of the likelihood. Whenever this happens, moreover, we can rotate the picture so that their centers are aligned vertically, and then the linear combination of θ_1 and θ_2 represented by the horizontal axis, which might be a feature of interest, will have a posterior that is centered to one side of the vertical alignment, as in the figure. So we can expect in general that the posterior will fail to be a compromise between the prior and the likelihood for at least some features of the parameter. When there are only a few parameters, we might anticipate this phenomenon and deal with it in some way, but it is increasingly difficult in high dimensions to see whether the phenomenon affects a particular feature of interest. This problem has been studied in detail by Min-ge Xie and his colleagues; see [243] and [244], pages 27ff and the discussion with Christian Robert on pages 55, 74–75.

These difficulties have become increasingly important in practice as well as in theory. Fields as disparate as medicine and macroeconomics now work with parametric models in which the dimensionality of the parameter space is orders of magnitude greater than the number of independent observations, and for such models prior probabilities dominate the analysis in ways not easily understood. Paul Romer, chief economist at the World Bank, has recently argued that this now happens routinely in the best respected work in macroeconomics [181].

The failure of the principles of stable estimation and compromise has left 21st century statistical theory in a quandary. This quandary can be seen as a crisis of Bayesianism, but I believe that it goes deeper, bringing into question not only the meaningfulness of a Bayesian prior for a Fisherian model with a large number of parameters but also the meaningfulness of such models themselves. We never have evidence that justifies such complex models, and we should consider probabilistic analyses that begin with fewer bets, bets that we do have some reason to trust.

2.5 The fiducial revival

The problems just discussed can be summarized by saying first that in complex models with a multidimensional parameter θ, our prior distribution will either overwhelm or distort the message of the likelihood function for some features $h(\theta)$, and second that this problem has become increasingly important in practice. Statisticians have understood for over half a century that a prior that seems relatively unopinionated about a large number of individual parameters $\theta_1, \ldots, \theta_n$ will express strong opinions about some features $h(\theta)$,[26] but now that we are working with so many parameters, in models so complex that their interaction is not transparent, this theoretical problem has become a real problem.

To deal with the problem, several statistical theorists have proposed focusing in advance on a feature $h(\theta)$ of interest and seeking posterior distributions that have desired Bernoullian properties for that particular feature. This is hardly consistent with Bayesianism's subjectivist philosophy, and it has sometimes led to non-Bayesian procedures that are variants on fiducial or Dempster-Shafer

arguments.

The theoretical statisticians exploring this direction of thought have not reached consensus on principles and methods, and I cannot survey their research in detail here. But here are three lines of thought that have attracted attention:

- **Confidence distributions.** The oldest and most obvious approach, perhaps, is to seek a method that produces nested confidence intervals for $h(\theta)$ at all levels of confidence and then to interpret these nested intervals as a probability distribution.[27] This approach was suggested by Bradley Efron in 1993 [83] and has since been developed by a number of authors, most notably Tore Schweder and Nils L. Hjort [186, 187] and Kesar Singh, Regina Liu, Min-ge Xie and their collaborators [244].

- **Generalized fiducial inference.** In this approach, developed by Jan Hannig and his collaborators [118], one chooses a structural equation $x = G(u, \theta)$ adapted to the feature $h(\theta)$ of interest.[28] After the observation of x, a posterior distribution for θ is found using Dempster's rule of conditioning (a special case of Dempster's rule of combination); the problem of conditioning on a set of measure zero in the continuous case is handled by first discretizing and then taking a limit. The posterior has desired Bernoullian properties under widely applicable conditions.

- **Inferential modeling.** This approach, developed by Ryan Martin and Chuanhai Liu [154], is also inspired by Dempster-Shafer theory; see [155]. Like generalized fiducial inference, it begins by adopting a structural equation $x = G(u, \theta)$ that determines the parametric model, but it then weakens the probability distribution for u to a Dempster-Shafer belief function (i.e., a random subset in u's space of possible values) in such a way that the structural function can be inverted without using Demspterian conditioning to obtain a Dempster-Shafer belief function for θ that has desired Bernoullian properties for $h(\theta)$.

Inferential modeling produces Dempster-Shafer belief functions that may or may not be probability distributions. The posteriors produced by generalized fiducial inference and confidence distributions are probability distributions (a probability distribution on the entire parameter space in the first case, and a probability distribution just for $h(\theta)$ in the second case), but there may or may not exist genuinely prior (not depending on the data) distributions that will give them as Bayesian posteriors.[29]

For a review of other methods for ensuring that Dempster-Shafer belief functions have Bernoullian properties, see [72]. For a recent study of fiducial methods that does not address the issue of Bernoullian properties, see [221].

3 The fiducial principle

The English words *fiducial* and *confidence* both derive from the Latin *fidere*, meaning "to trust". The first definition of *fiducial* given by the Oxford English

Dictionary is the general and theological one: "of or pertaining to, or of the nature of, trust or reliance". One example from 1870: "The words ... appear to ... fasten on the Lord with a fiducial grip."

When is a probability fiducial? Leaving aside Fisher's various answers to this question,[30] let us say that a probability becomes fiducial when we decide to trust it even though the evidence for it is weaker than we would like or even though we have other evidence that it ignores.

Once we adopt this broad sense of *fiducial*, we must recognize that practically all probabilities are fiducial when we put them to use. We create probabilities from theory, from conjecture, or from experience of frequencies. But this evidence is never strong enough to fully justify a system of numerical probabilities, and there is always other evidence.[31] To use the probabilities in a meaningful way, we must proceed nevertheless, and this makes the probabilities fiducial. This is just as true for Bernoullian and Bayesian probabilities as it is for the fiducial probabilities that Fisher invented. In fact, it is glaringly true at the outset for any Fisherian model $\{P_\theta\}_{\theta\in\Theta}$, for we never have enough evidence to fully justify the infinitely precise probabilities given by such a model.

What does it mean to trust a probability or a system of probabilities? Probably the best way to make this question concrete is to rephrase it in betting terms. What does it mean to trust given odds or a given system of odds? There are two distinct answers to this question. One answer, advocated by de Finetti and many other subjective Bayesians, is that we are disposed to bet at the given odds.[32] A second answer is that we respect them; we believe that no gambling strategy we devise to take advantage of them will make us rich without undue risk. More precisely: no gambling strategy will multiply the capital it risks by a large factor.[33]

We can qualify in many ways the notion that we trust or continue to trust certain odds or systems of odds. When interpreting this trust as disposition to bet, we can limit the disposition to specific bets, specific situations, specific opponents, and specific times. When interpreting it in terms of skepticism about gambling strategies, we can limit the skepticism to specific strategies undertaken at specific times. The lesson we should draw from the failure of Fisher's fiducial argument is that such limitations are sometimes needed, and the newer fiducial methods discussed in Section 2.5 impose such limitations.

When we take a closer look at the fiducial character of Bernoullian and Bayesian methods of using probability, we see that they also limit our trust in probabilities in various ways.

3.1 Bernoullian estimation

If an event with probability p happens y times in n independent trials, and n is large, then we can expect y/n to be close to p with high probability. In fact, if we specify a non-zero distance from p and a high probability, then we can find a value of n such that y/n will be at least that close to p with at least that probability. This is Jacob Bernoulli's theorem, first published in 1713. It is justly celebrated.[34] As Aleksandr Aleksandrovich Chuprov wrote

to commemorate its two hundredth anniversary [169], "everywhere the logic of inference rests in the final account on the theorem of Jacob Bernoulli."

Here is a slightly more formal statement of Bernoulli's theorem: For any $\epsilon > 0$ and any $\delta > 0$, we can find n large enough that the event

$$\left| \frac{y}{n} - p \right| \leq \epsilon \tag{5}$$

has probability at least $1-\delta$. This has many generalizations, all of which say that under certain conditions certain quantities can be estimated with high accuracy and high confidence. Chuprov's sweeping statement refers to the importance of these generalizations together with the original theorem.

The assertion (5) is uncontroversial when it is made before the trials, when we know n but not y. Should our subsequent knowledge of y change our probability for (5)? Do we know why and how we gained knowledge of y? Could the process that brought us this information be influenced by the process that determined p? Is it even possible that someone disclosed this information to us in order to mislead us about p? Use of Bernoulli's theorem in any particular case is legitimatized by the judgement that the additional information (including the value of y and the very fact that we have learned it) is not materially relevant to our use of the high probability for (5). This is a fiducial judgement. Similar judgements are required when we use the many generalizations and applications of Bernoulli's theorem.

Abraham De Moivre improved on Bernoulli's crude calculations by finding an estimate of the probability that the difference (5) will be within given bounds for a given n.[35] The logic for using this probability in the estimation of p was explained by Cournot in 1843. Here is the explanation, translated from the French but retaining Cournot's symbols: p for the probability we are estimating, n for the number of times the event happens in m trials (so that the frequency is n/m rather than y/n) and P for the probability before the trials are observed that $\left|p - \frac{n}{m}\right| \leq l$:

> As we have explained, the probability P has an objective value. It measures in effect the probability of error that we incur when we declare that the difference $\left|p - \frac{n}{m}\right|$ falls between the limits $\pm l$. Even if, for unknown reasons, certain values of p are able to appear more often than others in the ill-defined multitude of phenomena to which statistical observations can be applied, the number of true judgements that we will produce by declaring with probability P that the difference $\left|p - \frac{n}{m}\right|$ falls between the limits $\pm l$ will be to the number of mistaken judgements approximately in the ratio of P to $1 - \mathrm{P}$, provided that we make a large enough number of judgements that chance anomalies more or less cancel each other out.[36]

Here Cournot envisages a sequence of problems in which the unknown p varies but we select an interval of the same probability P each time. (The numbers m and n may also vary, as will the length of the interval.) By another application of Bernoulli's theorem, we will be right P of the time. The fiducial judgement

here is that we are content with this – that we are content that the probability P will resist anyone who bets against it before seeing the outcomes.

Cournot's logic merely elaborated the reasoning of Laplace and Gauss before him, when they derived what we would now call large-sample non-Bayesian confidence limits (see [116], Chapter 8). It also seems fair to say that this logic was the implicit foundation for the exposition of least squares in most manuals on statistics and error theory for the following century (see for example Bowley [19] and Palmer [171]), even if some expositions (Poincaré's text [174], for example) evoked inverse probability to justify least squares. Neyman made the same argument as Cournot had made when he called the use of confidence intervals *inductive behavior* in 1957 [165].[37]

Is P still a probability after the observations are made? Neyman said no. As he explained to de Finetti in 1939,

> My expression "confidence coefficient" designates the *value* of the probability that an estimation is correct, a value chosen arbitrarily in advance; so this expression is not a synonym for the term "probability".[38]

The fiducial principle can free us, however, from debates about whether a particular number used in a particular way is or is not a probability. We can trust or continue to trust a probability in different ways, and we need not subscribe to a doctrine, be it de Finetti's or Fisher's, that prescribes a panoply of ways we must trust it in order to continue to call it a probability.

3.2 Bayesian estimation

Bayesian estimation is usually explained in a formal way. Bayes's theorem is deduced from the definition of conditional probability and used in a Fisherian model in which the probabilities appear as conditional probabilities given the parameters. Attention is then directed to the choice of initial probabilities for the parameters, and the philosophical discussion revolves around the subjectivity of this choice.

The subjectively chosen initial or prior probabilities are evidently fiducial, for they are necessarily based on scanty evidence. Even "objective Bayesians", who believe that given evidence determines probabilities objectively, generally concede that they lack the resources to calculate or otherwise determine those probabilities precisely. As Alan Turing put it ([224], Section 1.3; [248]), "When the whole evidence about some event is taken into account it may be extremely difficult to estimate the probability of the event, even very approximately, and it may be better to form an estimate based on a part of the evidence, so that the probability may be more easily calculated."

Moreover, the use of Bayes's theorem adds further fiducial judgements. In Thomas Bayes's time, there was no such thing as conditional probability – no such general concept, no formal definition, and certainly no notation for it. But earlier writers had considered events that happen or fail in sequence, and they had considered how probabilities for later events change as earlier ones happen.

Abraham De Moivre, for example, considered an event A and a later event B and showed that the probability of B after A happens, for which I will write $P(B|A)$,[39] should satisfy

$$P(A\&B) = P(A)P(B|A), \tag{6}$$

where $P(A)$ and $P(A\&B)$ are the initial probabilities for A and $A\&B$, respectively. The equality (6) has long been called the *rule of compound probability*. It implies, of course, that

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

when $P(A) > 0$. De Moivre's argument for the rule of compound probability was based on the betting definition (or the game-theoretic definition, as we can now call it) of probability: the probability of an event is the amount you must risk to end up with one monetary unit if the event happens.[40] To turn $P(A)P(B|A)$ into one monetary unit if $A\&B$ happens, you first bet it all on A; this gives you $P(B|A)$ if A does happen, in which case you bet this on B.

In his famous essay on probability, published posthumously in 1763, Bayes repeated De Moivre's proposition and proof; this was his third proposition. But he also tried to prove an analogous result backwards in time: if you learn B has happened without knowing whether the earlier event A has happened, you should change your probability for A from $P(A)$ to

$$\frac{P(A\&B)}{P(B)}. \tag{7}$$

This is the fifth proposition in Bayes's essay, but his proof was hardly a proof. He imagined a sequence $(A_1, B_1), (A_2, B_2), \ldots$ of events ordered in time and posited that we will be told nothing about which ones happen until the first B happens. Then we will be told that this B has happened, and we will bet on the A that is paired with it. Thus we know in advance that we will be told B and will have no other information. The argument for changing from $P(A)$ to the ratio (7) is then convincing. But this does not establish that the change makes sense in other cases, where we may have other information, or we may not have known in advance what we would be told and when, so that the very fact that are told about B without being told about A is itself information [193]. To use Bayes's fifth proposition, we must make the fiducial judgement that this additional information is irrelevant. We must decide, as Bruno de Finetti explained centuries later, that this additional information does not change our attitude towards certain bets.[41]

Were we to accept Bayes's fifth proposition, and were we then to adopt uniform prior probabilities for an unknown prior probability p, then we could derive Bayes's formula for Bernoulli's problem of estimating p from y happenings in n trials:

$$\text{posterior probability that } a \le p \le b = \frac{\int_a^b p^y(1-p)^{n-y}dp}{\int_0^1 p^y(1-p)^{n-y}dp}. \tag{8}$$

Bayes's friend Richard Price, who published Bayes's essay after Bayes's death, explains in his introduction to the essay that Bayes had written this argument out but had feared that his readers would not find it convincing and had therefore used a different argument, an argument involving rolling balls on a rectangular table, and bolstered this argument with a scholium. This argument did not appeal to Bayes's fifth proposition.

The table's two dimensions are not needed, and we can explain the argument more quickly in one dimension, as Morgan Crofton did in the article on probability in the *Encyclopædia Britannica* in 1885 [39]. The question "will not be altered" Crofton opined, if we suppose that whether the event happens or not on each trial is determined by whether a point chosen at random on a line segment falls to the left or the right of a particular unknown point. Suppose, for simplicity, that the segment is the unit interval $[0, 1]$; the event happens if the point falls to the left of p, fails if it falls to the right of p; thus it happens each time with probability p. The point p itself is also chosen at random – i.e., from the uniform distribution on $[0, 1]$. So all we know of p is that it is the $(y + 1)$st in order of $n + 1$ points chosen at random in A. The formula (8) follows. The fiducial judgement here the assumption that the random choice of the point p is independent of the statistical evidence y – independent of the random choices of the n other points on the line. This replaces the equally fiducial fifth proposition.

In his scholium, Bayes pointed out that y has $n + 1$ possible values, namely $0, 1, \ldots, n$, and that his billiard table experiment, considered before any throws are made, gives equal probabilities to these $n + 1$ values. The reasonableness of this result, he contended, vindicated his method.

In his introduction, Price asserts that it was the uniform prior probabilities for p that Bayes feared might be unpersuasive.[42] This uniform distribution is still present and hardly disguised, however, in the billiard-table argument. So it seems reasonable to ask if it might instead have been the fiducial argument for the fifth proposition that worried Bayes. Assumptions of independence were familiar and easily accepted even in Bayes's time, and so the assumption that p was chosen independently of the other points might have seemed more persuasive.

The first person to explain the limitations of Bayes's rule clearly may have been Antoine Augustin Cournot, in his 1843 book, *Exposition de la théorie des chances et des probabilités*. He summarized his analysis as follows:

> Bayes's rule ... has no utility aside from leading to the fixing of bets under a certain hypothesis about what the arbiter knows and does not know. It leads to an unfair fixing if the arbiter knows more than we suppose about the real conditions of the random trial.[43]

The fiducial principle allows us to say this in a more positive way: we should continue to trust the betting rate only if we make the judgement that other information, information other than B's happening and the information that went into fixing $P(B)$ and $P(A\&B)$, is irrelevant.

3.3 Dempster-Shafer belief functions

The arguments by Bayes and Crofton that we just reviewed can be placed within Dempster-Shafer theory and generalized in various ways. Dempster's first article on the theory included a generalization of the Crofton argument in which we do not put prior probabilities on p and hence obtain only upper and lower posterior probabilities for it [59].[44] In [68], Dempster explained how the simple fiducial example discussed on page 10 above fits into Dempster-Shafer theory, where it generalizes to a treatment of the Kalman filter.

The central idea of Dempster-Shafer theory is what I call *Dempster's rule of combination*. This rule tells us how to combine beliefs (upper and lower probabilities) based on independent bodies of evidence. Here (as in the case of Bayes's billiard table), the word *independent* signals a fiducial judgement. We decide that each body of evidence does not materially change certain judgements based on the other body of evidence. As Dempster occasionally put the matter to me in the 1970s, we "continue to believe". As I now prefer to say, we continue to trust that certain bets will not succeed spectacularly. Over the years, critics of Dempster-Shafer have pointed to examples where we do not want to make this judgement, but that there are such examples only confirms that the judgement is needed. Bayesian arguments are in the same boat.[45]

The fiducial principle is also at play in Dempster-Shafer and Bayesian analyses in financial auditing, intelligence, and other domains where little of the evidence we want to combine is statistical. In these cases, we construct belief functions or probability distributions by drawing an analogy between games of chance and other setups where numerical probabilities are strongly trusted and murkier situations to which we decide to extend that trust. See [192, 205].

3.4 Imprecise and game-theoretic probability

As I have noted (page 18), we can decide to continue to trust only certain probabilities rather an entire initial probability distribution. We do this when we "condition" using the formula (7), for this amounts to continuing to trust certain conditional bets while no longer trusting the initial unconditional bets. The same move is involved (if renormalization is required) in the more general case of Dempster's rule of combination. Generalized fiducial inference, discussed in Section 2.5, also involve this move.[46]

If we anticipate that we might retain only some probabilities, it is reasonable to ask whether some of those that we will not retain can be identified at the outset and removed from the initial model, thus making this model simpler and perhaps more plausible as a representation of actual evidence. This may take us outside the Fisherian framework and into the realm of imprecise and game-theoretic probability [5, 206]. For an application of the fiducial idea to the theory of imprecise probability, see [204]. For a yet more general picture in which different probability judgements are trusted to different degrees, see [109].

The theories of imprecise probabilities and game-theoretic probability both

begin with a betting interpretation of probability but differ in how they trust given odds. Following Walley [237], most authors working with imprecise probabilities adopt de Finetti's view that subjective probabilities are dispositions to bet or otherwise act. In the game-theoretic picture [206], trust in a system of odds is interpreted as a judgement that these odds will resist gambling strategies: they will not allow you to multiply the capital you risk by a large factor. This is the game-theoretic version of Cournot's principle (see Section 5 below).

4 Poisson's principle

Siméon Denis Poisson (1781–1840) was Laplace's successor as the leader of French mathematics [24]. We can trace back to his work in the 1830s the principle that probabilistic prediction is possible even when probabilities vary.[47]

In 1835, Poisson enthusiastically announced what he saw as a great empirical discovery:

> Things of every nature are subject to a universal law that we may call *the law of large numbers*. It consists in the fact that if you observe very considerable numbers of events of the same nature, depending on causes that vary irregularly, sometimes in one direction and sometimes in another, without tending in any particular direction, you will find a nearly constant ratio between these numbers.[48]

Poisson explained this empirical stability by generalizing Bernoulli's theorem. He showed that with high probability, counts and averages will be stable over time even if the probabilities and expected values vary.

Poisson's contemporaries found the complexity of his picture confusing. If there are probabilities for how the probabilities vary, then Bernoulli's theorem, applied to the mean probability, is theory enough.[49] But they took up his insight in various ways. In 1846, for example, the Russian mathematician Pafnuty Chebyshev (1821–1894) proved a generalization of Bernoulli's theorem in which the probabilities vary [30]. Many other generalizations followed.

4.1 Beyond the random sample

For simplicity, let us first consider the case where we consider the frequency of some event in successive trials (rather than the average of some variable quantity), but where the event's probability may change. We may suppose that the trials are tosses of a coin. Suppose there are n successive tosses. Set

$$x_n := \begin{cases} 1 & \text{if the } n\text{th toss comes up heads} \\ 0 & \text{if the } n\text{th toss comes up tails,} \end{cases}$$

so that $\sum_{i=1}^{n} x_i/n$ is the frequency of heads in the n tosses. Here are three successively more general versions of the law of large numbers, ϵ and δ being arbitrarily small positive numbers.

Version 1 (Bernoulli). Suppose the tosses are independent and the probability p of heads is the same each time. Then we can find a value of n sufficiently large that

$$\left| \frac{\sum_{i=1}^{n} x_i}{n} - p \right| \leq \epsilon \tag{9}$$

with probability at least $1 - \delta$.

Version 2 (Chebyshev). Suppose the tosses are independent and the probability of heads on the ith toss is p_i. Then we can find a value of n sufficiently large that

$$\left| \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} p_i}{n} \right| \leq \epsilon. \tag{10}$$

with probability at least $1 - \delta$.

Version 3 (Lévy). Suppose P is a probability distribution for $x_1, \ldots, x_n$. Then we can find a value of n sufficiently large that

$$P\left(\left| \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} E(x_i | x_1, \ldots, x_{i-1})}{n} \right| \leq \epsilon \right) \geq 1 - \delta, \tag{11}$$

where $E(x_i | x_1, \ldots, x_{i-1})$, the expected value under P of x_i given the values of $x_1, \ldots, x_{i-1}$, is also the probability that $x_i = 1$ given $x_1, \ldots, x_{i-1}$.

In each version, the conclusion of the theorem is that the frequency of heads will approximate, with very high probability, a probability or an average probability. In Version 1, the frequency approximates the probability p. In Versions 2 and 3, it approximates an average probability. Version 3, the martingale law of large numbers, began to emerge only with the work of the Russian mathematician Sergei Bernstein in the 1920s and was first clearly understood by the French mathematician Paul Lévy in the 1930s.[50] British and American mathematical statisticians began to think in terms of Versions 2 and 3 only beginning in the 1940s, as they more fully absorbed continental work on mathematical probability as a result of the influx of mathematicians fleeing Hitler.[51]

4.2 Beyond frequencies

Bernoulli's, Chebyshev's and Lévy's laws of large numbers for coin tossing all generalize to the case where the random variables $x_1, \ldots, x_n$ are not necessarily binary but satisfy certain regularity conditions. The ratio $\sum_{i=1}^{n} x_i / n$ is then an average, not necessarily a frequency; p in (9) is x's mean; p_i in (10) is x_i's mean. The conditional expected value in (11) is no longer necessarily a conditional probability.

Poisson's principle, as I have formulated it, says that these laws of large numbers are predictions, even though they are not statements about frequencies. The prediction

$$\left| \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} E(x_i | x_1, \ldots, x_{i-1})}{n} \right| \leq \epsilon \tag{12}$$

in (11), when the x_i are not binary, is a case in point. It does not equate a probability or even an average probability with a frequency.

Poisson's principle is now a commonplace. Markov processes, martingales, time-series models, and a plethora of other stochastic processes have been major topics of statistical research for more than half a century. But our ways of talking have sometimes lagged behind, remaining in Fisher's picture of a random sample from a hypothetical population. The persistence of the word *frequentist* is one example of this lag.

In 1960, in the *Journal of the American Statistical Association* [166], Jerzy Neyman announced that stochastic processes had superseded independent trials in all branches of science. He wrote:

> The fourth period in the history of indeterminism, currently in full swing, the period of "dynamic indeterminism," is characterized by the search for evolutionary chance mechanisms capable of explaining the various frequencies observed in the development of the phenomena studied. The chance mechanism of carcinogenesis and the chance mechanism behind the varying properties of the comets in the Solar System exemplify the subjects of dynamic indeterministic studies.

Here he was evidently using *frequencies* in a broad and even metaphorical way, to refer not merely to frequencies on repeated trials but to averages of various kinds.

The law of large numbers is further generalized game-theoretically in [206], from the setting where a probability distribution for the whole sequence of variables is offered at the outset to the case where possibly more limited bets are offered on x_i after $x_1, \ldots, x_{i-1}$ are announced. For example, you may be offered x_i at the price m_i. Assuming for example that the x_i and m_i are all uniformly bounded in absolute value, we can show that for n sufficiently large,

$$\overline{P}\left(\left|\frac{\sum_{i=1}^n x_i}{n} - \frac{\sum_{i=1}^n m_i}{n}\right| \le \epsilon\right) \le 1 - \delta, \tag{13}$$

where $\overline{P}(A)$, the upper probability of an event A, is by definition the amount of money you must risk in order to get one monetary unit if A happens.

5 Cournot's principle

To put Poisson's principle to work, we must acknowledge how a probabilistic theory makes a prediction: it predicts an event by giving it very high probability. As Abraham Wald said in a lecture at Notre Dame in 1941 [235], probability theory can be applied to real phenomena by translating the theoretical statement "the event E has a probability near to one" into "it is practically certain that the event E will occur in a single trial."

We did not need Wald to tell us this. As soon as we saw the probability statement (11), we understood that the stochastic process represented by P was

predicting the event (12). But this needs to be stated explicitly. Cournot did so, and he also pointed out that this is the *only* way that probability relates to phenomena [207].

If we agree with Chuprov that Bernoullian statistics rests on Bernoulli's theorem and its generalizations, then we must also recognize that Cournot's principle is part of that foundation. Chuprov and his student Oscar Anderson called it *Cournot's bridge*, because it connects the probability statement (e.g., Bernoulli's theorem) to the event it predicts (e.g., the empirically observed law of large numbers) [156, 157]. It was the French mathematician Maurice Fréchet who first called this bridge *Cournot's principle*.[52]

In addition to providing part of the foundation of Bernoullian statistics, Cournot's principle also helps bring Bernoullian and Bayesian statistics together, because most statisticians who call themselves Bayesian also believe in model checking. In the end, a Bayesian model is of little use in practice unless its predictions are consistent, in the large, with what we observe. For Bayesian testimony on this point, see George E. P. Box's classic defense of significance testing [20]. See also [101, 189, 203].

Most continental mathematicians who studied mathematical probability in the first half of the 20th century subscribed to Cournot's principle in one way or another. In addition to Wald, salient examples include Evgeny Slutsky, Paul Lévy, Emile Borel, Andrei Kolmogorov, Abraham Wald, and Trygve Haavelmo [156, 208, 198, 202]. Like Chuprov, these mathematicians saw Bernoulli's theorem and its generalizations as fundamental to probability, but they also saw that only one of the probabilities in Bernoulli's theorem is being approximately equated with a frequency. The probability p in (5) is equal for practical purposes to the frequency y/n, but the probability that p is within ϵ of y/n is not a frequency. As Cournot explained (see page 19 above), we can interpret it as a frequency if we want; we imagine that the whole experiment involving n trials is itself repeated many times. Again applying Bernoulli's theorem, we see that with very high probability the frequency with which $p - y/n \leq \epsilon$ happens in this imaginary superexperiment will be near one. But again we have an uninterpreted probability, and to give it a frequency interpretation we need a yet huger imaginary superexperiment. Here looms, in the words of R. A. Fisher, "a perpetual regression defining probabilities in terms of probabilities in terms of probabilities" ([96], page 266). We will have an uninterpreted probability forever unless we terminate the regression by applying Cournot's principle.

The continental mathematicians also saw that the law of large numbers is far from being the only prediction that can be checked in order to test a probabilistic hypothesis. Consider, for example, the law of the iterated logarithm, which concerns the rate at which a frequency will converge in repeated independent trials of an event [126]. This law is of little interest to statisticians, because the number of observations needed to test it is impossibly large, but there are related predictions that can be tested. One such prediction, emphasized by Jean Ville, is that the frequency will at least oscillate around the probability; it will not converge to it from above or from below [227].

5.1 Objections to the principle

Cournot's principle is sometimes criticized for its vagueness. A probability close to one is a prediction, but how close to one? As Wald explained, such vagueness is always associated with the application of theory to real phenomena:

> The purpose of statistics, like that of geometry or physics, is to describe certain real phenomena. The objects of the real world can never be described in such a complete and exact way that they could form the basis of an exact theory. We have to replace them by some idealized objects, defined explicitly or implicitly by a system of axioms. For instance, in geometry we define the basic notions "point," "straight line," and "plane" implicitly by a system of axioms. They take the place of empirical points, straight lines, and planes which are not capable of definition. In order to apply the theory to real phenomena, we need some rules for establishing the correspondence between the idealized objects of the theory and those of the real world. These rules will always be somewhat vague and can never form part of the theory itself.

On the other hand, we can give guidelines, depending on context and purpose. Emile Borel, who called Cournot's principle "the only law of chance" [18], suggested that a probability of 10^{-6} is negligible on a human scale, a probability of 10^{-15} on a terrestrial scale, a probability of 10^{-50} on a cosmic scale, and a probability of 10^{-1000} on a universal scale ([17] pages 6–7).

Another common objection is that what happens always has small probability. A lottery always has a winning ticket. This overlooks the role of the statistician or scientist, who chooses the prediction in advance.[53] Injecting a scientist into the picture might seem to threaten the objectivity of the probability model, but in practice only a limited number of predictions are important [23, 234]. Even in theory we can only make a countable number of predictions, which could be combined into a single prediction were it computable [15].

Bruno de Finetti rejected Cournot's principle.[54] In a note to Maurice Fréchet in 1955 [55], he wrote as follows:[55]

> The definition of subjective probability is based on the behavior of the person who assesses it: it comes down to measuring the sacrifices the person thinks it reasonable to accept in order to escape from the risk of some harm that would accompany the event considered (insurance premium, betting rate, etc.). In particular, saying that the probability is small indicates that the person judges the risk to be negligible, that is to say, that he acts more or less as if the event were impossible. If this is not a principle, it is because it is by definition a synonym, a tautology, a banality.

This passage might give the impression that de Finetti would agree that a very small probability authorizes a prediction. But this he consistently denied. In his uncompromising subjective view, a probability is a always forecast, never a prediction.[56]

5.2 The game-theoretic form of the principle

We predict using a probabilistic theory by singling out an event E to which it gives small probability and predicting that E will not happen. This is one of the simplest ways of stating Cournot's principle. We can make the statement more concrete by interpreting the theory's probability for E, say α, as an offer to bet. If we take the bet, betting α on E, and E does happen, then we will have turned the amount α we have risked into the amount 1; we will have multiplied the capital we risked by the large factor $1/\alpha$. So predicting that E will not happen is tantamount to predicting that we will not multiply our capital by a large factor.

The game-theoretic form of Cournot's principle can be stated in a somewhat more general and flexible form: *the objective content of a system of probabilities lies in the prediction that a strategy for betting at the corresponding odds will not multiply the capital risked by a large factor*. This coheres with the game-theoretic form of the fiducial principle, as stated on page 18: to use probabilities, we must decide to believe that no strategy we devise to bet at the corresponding odds will multiply the capital it risks by a large factor.

One advantage of the game-theoretic formulation is that it makes more salient the requirement that the event being predicted be selected in advance of observing whether it happens. No one will take a bet after the fact.

As noted on page 18, we can use the fiducial principle flexibly by deciding only to trust certain probabilities and only in certain situations. This flexibility is particularly salient and far-reaching when we use the game-theoretic formulation. Instead of predicting that no strategy for using given odds will multiply the capital risked by a large factor, we can make this prediction for a particular strategy or for some small class of particularly simple strategies.

The generalization from predicting the failure of particular events to predicting the futility of particular strategies is particularly consequential when we use Cournot's principle for testing. In this case, the choice of a particular event (or *critical region*, as it is called in the Neyman-Pearson theory of testing) also fixes the significance level α, which measures the strength of the test and hence the level of negative evidence if the test rejects. A gambling strategy, on the other hand, can measure the level of negative evidence more flexibly, by how large a factor the capital is multiplied. This measure has a legitimacy not shared by p-values. When we use p-values, we are specifying in advance a test statistic on which to bet, but we are not specifying the bet in advance. Moreover, idea of testing by a gambling strategy is applicable when we are dealing with a forecasting system, an actual forecaster, or a theory that produces only limited betting odds, perhaps sequentially with feedback, whereas the idea of testing by selecting an event of small probability requires that the theory or forecaster provide a comprehensive probability distribution in advance so that a critical region can be selected.

See Appendix II for additional information on the game-theoretic understanding of probability.

6 Conclusion

In practice, both Bernoullian and Bayesian statistics rely on fiducial judgements. Bernoullian statistics relies on judgements, made in particular cases, that predictions in which we are confident before certain observations still merit our confidence after. Bayesian statistics relies on similar judgements, applied to conditional predictions.

Poisson's principle clarifies the role of frequencies. Bernoullian and Bayesian analyses make predictions about averages and about other events, not merely about frequencies.

Cournot's principle tells us how a probabilistic analysis, Bernoullian or Bayesian, makes a prediction: it assigns the predicted event a probability close to one. This can be put in betting terms. The probability close to one implies very favorable odds for a bet against the event, odds that would multiply the capital you risk by a large factor if the event fails; the prediction is that the bet will not succeed.

The three principles unify probability while validating its diversity. They are used by fiducial, Dempster-Shafer, and imprecise-probability analyses just as they are used by Bernoullian and Bayesian analyses. This lends legitimacy to these less classical approaches and may open the way to even leaner paradigms of probabilistic analysis and prediction. Bets we choose to trust may yield interesting predictions even if they are too sparse to define random points or random subsets or to satisfy the axioms of imprecise probability.

7 Appendix I: *Bayesian*, *Bernoullian*, etc.

How did the names Bayesian, fiducial, and frequentist arise? What other names have been used for the Bayesian and Bernoullian schools of thought?

7.1 Bayesian

So far as we know, *Bayesian* has been used in English to refer to the work of Thomas Bayes only beginning in the middle of the 20th century. In the second half of the 19th century and the first half of the 20th, we find only *Bayes's* or *Bayes'*, as in *Bayes's rule*, *Bayes's formula* and *Bayes's theorem*. We similarly see only the possessive form in French during this period: *régle de Bayes*, not *règle bayesienne*.

We do see the adjectival form very early in German. The German translation of Cournot's book on probability, which appeared in 1849, translated Cournot's *règle de Bayes* as *Bayes'sche Regel*. The adjective endured. Emmanuel Czuber used *Bayessche* in his history of probability ([42] 1900) and in the multiple editions of his authoritative probability textbook ([43] 1903). In the German edition of Andrei Markov's textbook, published in 1912 [153], we find both *Formel von Bayes* and *Bayesschen Formel*. In his book on the philosophy of

probability ([44] 1923), Czuber applied the adjective *Bayessche* to the nouns *Theorem*, *Satz*, *Formel*, *Regel*, *Ansatz*, and *Schlussweise*.

This difference between German practice on the one hand and French and English on the other was not merely a matter of grammar or literary style. The English readily turned other prominent names into adjectives in the 19th century; witness *Newtonian*, *Kantian*, and *Laplacean*. The role of Laplace is surely the crux of the matter. He, rather than Bayes, developed the statistical methodology that we now call Bayesian, for Bayes studied only what we now call the binomial case. Yet it makes little sense to call the methodology Laplacean, for inverse probability was but one of the probability methods Laplace developed.[57] The English solved this problem by adopting the term *inverse probability*, which first appears in print in work by Augustus de Morgan in the 1830s, with reference both to Bayes's problem (finding an inverse or converse to Bernoulli's theorem) and Bayes's and Laplace's solution of the problem [81].[58] The French, who became remarkably disinterested in and even hostile towards Laplace's work on probability during the second half of the 19th century [28, 158], continued to use Cournot's name *règle de Bayes* and similar phrases. The Germans, as we have noted, continued to use *Bayessche*.

The influx of German-speaking mathematicians into the United States and Britain before, during, and after World War II surely brought German ways of speaking with it. In any case, *Bayesian* begins to appear in print in English around 1950. The first appearance I have seen came in 1948, in an article by Charles P. Winsor, then working in biostatistics at Johns Hopkins [241]. Reviewing a discussion of binomial estimation that had taken place in the *Educational Times* in the 1880s, Winsor uses the phrases *Bayesian argument* and *Bayesian assumption*. The next appearance is in 1950, in prefaces R. A. Fisher wrote for two of his earlier papers [93].[59] In 1951, L. J. Savage writes of "modern, or unBayesian, statistical theory" [183].

As Stephen Fienberg has documented, the adjective *Bayesian* became standard in the 1950s [86]. Those who began using it then included long-standing advocates of Bayes's rule such as I. J. Good, newly Bayesian statisticians such as Savage and Denis Lindley, and decision theorists in American business schools such as Harry V. Roberts and Robert Schlaifer. Good had learned probability by reading Keynes and Ramsey in the 1930s (see the preface to [107]) and had learned inverse probability by working with Turing in World War II. According to Fienberg, Good first used *Bayesian* in 1956, in a review in *Mathematical Reviews* of an article in French by de Finetti, where de Finetti had used *bayesien*; Good subsequently used *Bayesian* in an article published in 1958 [106]. Lindley was perhaps the first to use *Bayesian* extensively and systematically in print, in an article published in 1958 [145], and as Fienberg notes (page 17), Savage used the adjective in corresponding with Lindley about a draft of this article. In 1958, Erich Lehmann used *Bayesian derivation* in passing in an unpublished technical report [140]. In a symposium Savage led in London in 1959 (published only in 1962), the adjective was used by Savage, Peter Armitage, George Barnard, Maurice Bartlett (quoting Lindley), and David R. Cox. By 1960, Roberts could

write that "Bayesian statistics" was now a standard term ([179], page 26). By 1962, he could write about the sometimes-called "Bayesian revolution" ([180], page 202).

In the instances just cited, *Bayesian* was used as an adjective. The earliest instance of the word being used as a noun that I have located is by Savage in 1961; he writes ([184], page 577):

> We Bayesians believe that the dilemma to which the frequentist position has led, along a natural and understandable path, is insoluble and reflects what is no longer a tenable position about the concept of probability.

The dilemma to which he refers is simply the inability of frequentists (Bernoullian statisticians) to express their conclusions in the form of probabilities for hypotheses.

Counterparts for the newly coined English *Bayesian* and *Bayesianism* eventually came into use in other European languages. The first uses I have seen in Italian and French were by Bruno de Finetti; he used the adjective *bayesiano* in Italian in 1954 [53] and the adjective *bayesien* in French in 1955 ([54], the article reviewed by Good). The French adjective is now written more often as *bayésien*, in an attempt to better imitate the English pronunciation. The French noun *bayésienisme* came much later and is still rare. In German, the English noun *Bayesian* became *Bayesianer* and *Bayesianism* became *Bayesianismus*.

7.2 Bernoullian

In this article I have used the adjective *Bernoullian* to refer in general to non-Bayesian methods of statistical testing and estimation that are now often called *frequentist*. This usage is not standard but has a reasonable pedigree, going back at least to Francis Edgeworth:[60]

- Edgeworth used *Bernoullian* with this meaning in 1918, contrasting "the *direct* problem associated with the name of Bernoulli" with "the *inverse* problem associated with the name of Bayes" [79].

- Richard von Mises made a similar remark in German in 1919 ([229], page 5): "Man kann die beiden großen Problemgruppen ... als den Bernoullischen und den Bayesschen Ideenkreis charakterisieren." In English: "We can call the two large groups of problems the Bernoullian and Bayesian circles of ideas."

- Arthur Dempster advocated the usage in 1966 [59]. In 1968 [62], in a review of three volumes of collected papers by Neyman and Pearson, Dempster wrote

  > Neyman and Pearson rode roughshod over the elaborate but shaky logical structure of Fisher, and started a movement which pushed the Bernoullian approach to a high-water mark from

> which, I believe, it is now returning to a more normal equilibrium with the Bayesian view.

- Ian Hacking used the term several times in his 1990 book, *The Taming of Chance* [110]. Writing about Poisson's interest in changes in the chance of conviction by a jury, he wrote (page 97):

  > Laplace had two ways in which to address such questions. One is Bernoullian, and attends to relative frequencies; the other is Bayesian, and is usually now interpreted in terms of degrees of belief. Laplace almost invited his readers not to notice the difference.

The adjective *Bernoullian* honors Jacob Bernoulli just as *Bayesian* honors Thomas Bayes, and in a parallel way. In both cases, the person honored dealt only with the estimation of an individual probability, but their approach has grown into a vast methodology. Unlike *frequentist*, moreover, *Bernoullian* does not suggest a naive equation of probability with frequency.

In addition to *frequentist*, Bernoullian statistics has also been called *objectivist*, *orthodox*, *classical*, and *sampling-theory*. I turn now to these names.

Classical

Although Edgeworth's use of *Bernoullian* in 1918 is notable, the need for such a name was widely felt only in the mid-twentieth century, when Bayes's rule was first widely seen as a general methodology rather than a particular method. The need was first felt by the Bayesians, who needed a name for their opponents. Savage's *objectivistic* and the occasionally used *non-Bayesian* were awkward, and *modern*, used by Savage before he considered himself a Bayesian, would no longer do. The adjectives *orthodox* and *classical* were better suited to the occasion, and both were common in the 1950s and 1960s. It is easy to find authors who used both adjectives, and others as well:

- I. J. Good used *orthodox statistical theory* in his 1950 book, *Probability and the Weighing of Evidence* [104]. In a 1956 book review, he used *orthodox* and *classical* in the same paragraph ([105], page 389). In a 1958 article, he used *classical objectivistic statistics* [106]

- Edwards, Lindman, and Savage systematically contrasted *classical* with *Bayesian* statistics in their 1963 article [82].

- Denis V. Lindley used *classical statistics* in a 1964 article [146]. In the preface to a 1965 book [147], he used *orthodox statistics*.

- John W. Pratt, in a 1965 article entitled "Bayesian interpretation of standard inference statements" [177], explained that by "standard" he was referring to methods developed in the "orthodox", "classical", "objective", "frequency" or "Neyman-Pearson" tradition or traditions.

What is *orthodox* or *classical* is of course very changeable; these adjectives often refer to whatever aspect of yesterday's practice the author wants to replace or extend. The vagaries of *classical statistics* in the 20th century are particularly striking.

- Since the 1920s, physicists have used *classical statistics* to refer to statistical predictions that have been corrected by quantum theory.
- The preface to a statistics textbook published in 1940 [173] contrasted *classical statistics* as developed by Karl Pearson and his school with newer techniques developed by R. A. Fisher.
- In 1943, Jacob Wolfowitz contrasted *classical statistics* with nonparametric methods [242]. Joseph L. Hodges and Erich Lehmann were still using *classical* in this way in 1961 [120].
- For many in the mid 20th century, the treatment of inverse probability by Bayes and Laplace was classical. In a 1942 article in Russian ([129], page 4), Andrei Kolmogorov called the use of Bayes' theorem the classical method (классический метод). In the chapter on confidence regions in his 1946 book [36], Harald Cramér wrote (page 507):

  > In the older literature of the subject, probability statements of this type were freely deduced by means of the famous *theorem of Bayes*, one of the typical problems treated in this way being the classical problem of *inverse probability* ...

- Some authors in the 1950s and 1960s used *classical statistics* for methods that assumed random sampling, as opposed to newer methods for stochastic processes or time series. Examples include Geoffrey H. Jowett in 1956 and 1957 [123, 124], Donald A. Darling in 1958 [46], and John W. Tukey in 1961 [223].
- In 1953, M. A. Girshick used *classical statistics* to refer to Neyman-Pearson hypothesis testing, contrasting it with the theory of making decisions under uncertainty [102].

Although Girshick came close, it seems reasonable to say that I. J. Good was the first to use *classical statistics* as a general name for Bernoullian as opposed to Bayesian methods. He did so repeatedly, beginning in the 1950s. Also influential was the use of the term by Robert Schlaifer and his Bayesian decision-theory group at the Harvard Business School. Arthur Dempster has mentioned to me that this group's use of *classical* surprised him when he encountered it in the late 1950s; for Dempster as for Cramér, inverse probability was classical, and Neyman-Pearson theory was the innovation. In the chapter entitled "The Classical Theory of Testing Hypotheses" in his 1959 book, *Probability and Statistics for Business Decisions* [185], Schlaifer made his case for the terminology (page 607):

> At least in the United States, the theory of these procedures ... is now "classical" in the literal sense of the word: it is expounded in virtually every course on statistics and is adhered to by the great majority of practicing statisticians.

One remarkable aspect of this use of the name *classical statistics* is that some proponents of the methods being called classical eventually adopted the term. It was used, for example, by Lucien Le Cam in 1964 [137] and by Jaroslav Hájek in 1967 [113]. Stephen Fienberg and John Aldrich have speculated that this embrace was influenced by Neyman's use of *classical probability* for the mathematics of probability that he had learned as a student in Poland. In Neyman's view, confidence intervals used classical probability to accomplish what Fisher was trying to do with his nonclassical fiducial probability [163].

Erich Lehmann continued to use *classical statistics* in the 21st century. In the first sentence of his *Fisher, Neyman, and the Creation of Classical Statistics*, posthumously published in 2011 [141], he wrote

> Classical statistical theory – hypothesis testing, estimation, and the design of experiments and sample surveys – is mainly the creation of two men: R. A. Fisher (1890–1962) and J. Neyman (1894–1981).

Frequentist

The thesis that probability should be equated with relative frequency was already being debated in the second half of the 19th century, but the word *frequentist* came into use much later. By all accounts, the word was first used in print by the Columbia University philosopher Ernest Nagel (1901–1985) in 1936 [160, 161]. Nagel used *frequentist* only as a noun; the Harvard University philosopher Donald Williams used it as an adjective and also used *frequentism* in 1945 [239]. Nagel and Williams used *frequentist* and *frequentism* to refer to a view about the meaning of probability, not to a statistical methodology. Thus *frequentism* was synonymous for them with the already common term *frequency theory of probability*.

In the 1920s and 1930s, many mathematicians used *frequency theory* to refer more specifically to the framework of Richard von Mises, which specified conditions on a sequence under which probability might be identified with limiting frequency in the sequence [229]. This framework was cumbersome compared with the axiomatics advanced by Fréchet and Kolmogorov [128], and by the end of the 1930s mathematicians had decisively rejected it as a starting point for mathematical work [208].

The term *frequentist* was first used to refer to Bernoullian statistics only in 1949, by the statistician Maurice G. Kendall [125], and it was not widely used before the 1960s. Jerzy Neyman bears some responsibility for its subsequent popularity. As we have seen, he used *frequencies* to refer broadly to the regularities predicted by stochastic processes. In a philosophical article published in 1977 [168], he emphatically embraced the label *frequentist*.

In this paper, I have argued against continued use of *frequentism* to refer to Bernoullian statistics. It suggests a naive equation of probability with frequency that hardly does justice to the generations of mathematicians who have developed the topic By using it, Bernoullian statisticians have persuaded many philosophers that their viewpoint is shallow and incoherent [112, 111].

Sampling-theory

The earliest use I have seen of *sampling theory* as a general name for Bernoullian statistics is by Denis V. Lindley, in a discussion paper published by the Royal Statistical Society in 1968 [148]. There Lindley uses "orthodox sampling theory description", "classical sampling theory methods", "sampling theory approach", and simply "sampling theory".

In a article published in 1971 [67], Arthur Dempster used similar language. He wrote (page 58):

> I do not believe that either the Bayesian approach or the sampling distribution approach to unity is a total error, but I do find that subtle issues are involved which compromise parts of both schools, so that a mixed viewpoint becomes desirable. Specifically, one must reckon with the weaknesses of sampling distribution methods for estimation and of Bayesian methods for significance testing.

In 1972, in another discussion paper for the Society [151], Lindley and Adrian F. M. Smith used "orthodox, sampling-theory framework" and "sampling-theory school". The response was strikingly different from the response to Lindley three years earlier, in that most of the discussants, some Bayesians and some not, followed his lead by using the same or similar variations on *sampling theory*. These included J. A. Nelder, David R. Cox, R. L. Plackett, A. P. Dawid, and C. Chatfield. Even Oscar Kempthorne used "sampling-theory school", though with the quotation marks.

Lindley continued to use the term in a number of later publications, including his well known 1975 article "The future of statistics: a Bayesian 21st century" [149] and in a number of later publications (e.g., [150]). Two other prominent statisticians whose repeated use of the term has attracted notice are George E. P. Box, who considered himself a Bayesian, and David R. Cox, who does not.

- Box contrasted the "sampling theory approach" to the Bayesian approach in his 1973 book with George C. Tiao, *Bayesian Inference in Statistical Analysis* [21]. In his well known 1980 discussion paper at the Royal Statistical Society ([20] 1980), Box also contrasted Bayesian theory with "sampling inference" and "sampling theory", and again a number of discussants followed by using similar terms.

- In their 1974 textbook, *Theoretical Statistics* [35], Cox and David V. Hinkley described theirs as the "sampling theory approach to statistical inference". This approach, they explained, follows the *repeated sampling principle* (page 45):

> . . . statistical procedures are to be assessed by their behavior in hypothetical repetitions under the same conditions.

In his *Principles of Statistical Inference*, published in 2006 [34], Cox wrote (page 7):

> There are two broad approaches, called *frequentist* and *Bayesian*, respectively, both with variants. Alternatively, the former approach may be said to be based on *sampling theory* and an older term for the latter is that it uses *inverse probability*.

In my view, *sampling-theory statistics* is even more misleading than *frequentism*, because it ties us so firmly to Fisher's framework of independent, identically distributed observations. It suggests, and the repeated-sampling principle makes explicit, the doctrine that a stochastic process that runs only once can be understood only by imagining that it runs many times – a doctrine that we can recognize as fallacious once we understand Cournot's principle. The resulting confusion extends beyond statistical work to fields in physics that use probability, including statistical mechanics [103], quantum mechanics [22], and cosmology [212].

7.3 Fiducial

At the beginning of his 1873 essay on determinism [159], James Clerk Maxwell wrote that "we need some fiducial point or standard of reference, by which we may ascertain the direction in which we are drifting." Maxwell was alluding to the use of the adjective *fiducial* in surveying and astronomy, where it refers, according to the Oxford English Dictionary, to a line or point, etc., assumed as a fixed basis of comparison.

Fisher was evidently also referencing this meaning of the word when he called the probabilities he constructed from a pivot fiducial. In his initial example, the fixed point was the 95th percentile of the cumulative distribution function of the pivot. By continuing to believe the 95% probability statement – by trusting it, we obtain a 95% probability bound on the parameter.

The analogy with a true fixed point is imperfect. What Fisher was taking as fixed is fixed only by a fiducial judgement. But he brought the word *fiducial* into statistics in a permanent way. Rather than leave it to designate merely a failed argument, I propose to use it in a wider way relevant to nearly every application of statistics.

8 Appendix II: Game-theoretic probability

The game-theoretic framework for mathematical probability can be traced back to ideas advanced by Blaise Pascal in the 17th century. In their correspondence in 1654, Pascal and Pierre Fermat had competing methods for calculating how

stakes should be divided when a competition is cut short before either player has won enough games to win the entire stakes. Fermat used combinatorial reasoning, of a kind that had been understood in Europe since the 13th century [8], whereas Pascal used backwards induction to study how the value of each player's position (or *expectation*) changes as games are won and lost. Fermat's reasoning used the notion of equally possibility chances, whereas Pascal's reasoning, especially as elaborated later by Christian Huygens, relied only on the players' agreement to play on even terms and hence seems to apply to games of skill just as well as to games of chance [209].

The classical definition of probability was based on Fermat's notion of equally possible cases: the probability of an event is the ratio of the number of favorable cases to the total number of cases. Pascal's approach, which begins instead with odds at which the players have agreed to bet, leads to a different definition of probability: the probability of an event is the amount you must risk in order to get one monetary unit if the event happens. The two definitions connect with phenomena in different ways: Fermat connects through the notion of "equally possible," though this may strike modern sensibilities as mysterious. Pascal needs some other connection, and the natural one is Cournot's principle: we assume that strategies for betting at the corresponding odds will not allow you to multiply the money you risk by a large factor.

Mathematical probability is developed starting with Pascal's game-theoretic definition in my 2001 book with Vovk [206] and in a series of subsequent papers by several different authors, many published and many posted at `probabilityandfinance.com`. Here are some highlights of this work:

1. Concrete versions of the classical limit theorems of probability (the law of large numbers, the central limit theorem, and the law of the iterated logarithm) follow from applying Cournot's principle to relatively simple strategies [206].

2. Abstract versions of more abstract measure-theoretic results, beginning with Lévy's zero-one law, can be deduced from assumptions about available bets similar to the assumptions used in the theory of imprecise probabilities [232].

3. Simple gambling strategies can be used to adjust p-values to account for the fact that they fall short of tests with fixed significance levels [48].

4. Reasonable statistical tests can be represented as betting strategies, and by playing against these strategies a forecaster can make a series of forecasts that pass the tests, provided only that he is provided feedback [199]. This casts light on why adaptive or non-stationary forecasting is possible and casts doubt on the notion that it teaches us anything about the world.

5. In securities markets, the assumption that a speculator will not multiply capital risked by a large factor relative to a market index using certain simple strategies implies that that the paths of security prices will look like (possibly time-distorted) geometric Brownian motion and that the

market index will appreciate in proportion to its accumulated variance [233, 201]. This resolves the equity premium puzzle and explains why equity performance is related to apparent risk without making assumptions about investors' probabilities and utilities.

Notes

1. The development of Fisher's thinking leading up the 1922 article has been studied by John Aldrich [1] and Stephen Stigler [218]. David Hand [117] has celebrated the article's significance. Anders Hald and Stephen Stigler have studied the early history of maximum likelihood ([115, 217], Chapter 16), and Stigler has reviewed its development by and after Fisher [219]. See also Vladimir Vovk's algorithmic treatment of the method's efficiency [231].

   What mathematical form might be given to Fisher's intuitive notion of a random sample from an infinite population? Fisher sketched an answer in a prefatory note to a 1925 article, "Theory of statistical estimation" [91].

2. In 1810, Laplace used what we now call characteristic functions or Fourier transforms to perfect his method of approximating integrals involving very high powers, a method that he had first begun to develop nearly forty years earlier. His 1810 breakthrough resulted in Gaussian or normal integrals and various instances of what we now call the central limit theorem. However, as Hans Fischer has noted ([87], page 23), Laplace applied his method to particular problems (often concerning errors of observations) and never stated a general theorem corresponding to the central limit theorem in today's sense. This work by Laplace is discussed by most authors who have studied 19th-century mathematical statistics. In addition to Fischer, these include Marie-France Bru and Bernard Bru [28], Richard William Farebrother [85], Prakash Gorrochurn [108], Anders Hald [114, 116], and Stephen Stigler [215].

3. Francis Edgeworth preceded Fisher in developing an abstract estimation theory that assumed a known class of probability distributions indexed by multiple constants, especially in his 1908 and 1909 articles on the "genuine inverse method" [77, 78]; see [116], pages 71–72. But Fisher's gift for exposition quickly earned him a much wider readership than Edgeworth ever enjoyed.

4. Was Fisher's systematic use of *parameter* original? In 1976 [214], Stephen Stigler reported finding only a few isolated instances where Fisher's British predecessors had used the word to designate a constant in a probability law. We can also find a few analogous instances of *paramètre* in French: by Auguste Bravais in 1846 (cited by Edgeworth, as Stigler noted), by Jean-Baptiste Liagre in 1852 [144], and by Emile Dormoy in 1888 [73]. But the word was also used in very different ways by French and English students of probability. In Henri Poincaré's probability textbook, for example, *paramètre* is sometimes used for what we now call a random variable ([174], page 98 in the 1896 edition, page 121 in the 1912 edition).

   In 1915, on the other hand, in the first edition of his probability textbook [88], the Danish-American statistician Arne Fisher had already systematically used *parameter* in the same way that R. A. Fisher later used it. See, for example, the discussion of "the parameters of frequency curves" on page 185. Notable in this connection is R. A. Fisher's 1931 letter to Arne Fisher, in which he defends his failure to cite the work of Scandinavian predecessors ([10], pages 310–313).

5. As John Aldrich has pointed out [3], Fisher put multiple regression in its modern form in another 1922 article [89] by adding the (often fictional) assumption that the values of the independent variables are initially known.

   Judy Klein has masterfully described the importance of time series in statistics from the 17th to the 20th centuries [127].

6. Some of their reservations about Fisher are on display in the discussion following his 1935 presentation to the Royal Statistical Society [92].

7. Bernard Bru has documented the emergence of the theory of stochastic processes in the 1920s [25]. See also the discussion of the martingale law of large numbers in Section 4.

8. Laplace did not use the terms *direct* and *inverse*; they were introduced, in English, by Augustus De Morgan in the 1830s.

9. As Bernard Bru and Pierre Crépel showed in their painstaking study of Condorcet's unpublished work, Condorcet used the inverse principle in an unpublished manuscript written before the spring of 1771. (See [31], pages 247–263, especially footnote 19 on pages 256–260; see also [37], pages 288–289, and [28].) Laplace began his work on the principle later, in the pioneering article he read to the Academy in 1773 and published in 1774. Immediately after stating the inverse principle in this 1774 article, Laplace applied it to the same problem (an urn containing finite but unknown numbers of black and white tickets) that Condorcet had considered in his unpublished manuscript ([131], pages 29–30). From this and other circumstantial evidence, Bru and Crépel concluded that Laplace most likely learned the principle from Condorcet's manuscript.

   Prior to Crépel's and Bru's work, Stephen Stigler had conjectured, based partly on study of an unpublished paper on the theory of errors drafted by Laplace in 1772, that Laplace had persuaded himself of the principle by way of a fiducial argument ([215], pages 100–101; see also [216]). It was in the theory of errors, in any case, that Laplace found the principle to have the greatest importance.

10. The French mathematicians studying probability in the 1770s had not taken serious notice of Bayes's work. D'Alembert and Condorcet apparently first noticed Bayes's priority around 1780, after Bayes's literary executor Richard Price had entered into correspondence with Turgot [31, 45]. Laplace first mentioned Bayes in print in his *Essai philosophique sur les probabilités* in 1814 [133].

11. Laplace argued that the averaging involved in least-square calculations leads to a normally distributed error of estimation regardless of the probability law for individual errors, and he deduced from this that least squares minimizes the error of estimation as measured by expected absolute error. He also argued that individual errors, if themselves the result of many independent influences, will be approximately normally distributed, vindicating Gauss's earlier inverse argument for least squares based on the assumption of normality. Later, in 1823, Gauss showed that least-squares estimators have the least mean-squared error among all unbiased linear estimators, thus providing a direct (as opposed to inverse) argument for least squares even for small samples For details, see the general histories cited earlier [85, 108, 116, 215, 222].

    A general abstract account of the equivalence of direct and inverse probability when the form of the probability law of the observations is known, showing that the sampling distribution of the maximum likelihood estimators in a model with multiple parameters has the same approximate multivariate normal distribution as the posterior distribution obtained using a uniform prior distribution, was first given by Francis Edgeworth in 1908 and 1909 (see Hald [116], Chapter 10).

12. Hald provides a succinct overview of these developments in [116]. His comment about Laplace and Gauss moving away from inverse probability in practice is on page 73. On page 101, he cites a letter from Gauss to Friedrich Wilhelm Bessel in which Gauss characterizes inverse probability as metaphysical.

    Contemporary reports on the status of the inverse-probability debate early in the 20th century are provided by Emanuel Czuber ([41], pages 91–110) and Arne Fisher ([88], pages 55–56).

    Czuber's own work provides an example of the distance between probability theory and the theory of errors at the end of the 19th century. His authoritative book on probability [43] included a chapter on Bayes's theorem, but his book on errors of observations [40] did not mention it.

13. On one occasion, Laplace suggested that prior information could be taken into account by introducing fictional observations; See [215], page 136.

More generally, unequally possible cases can be made into equally possible cases by subdividing cases that are more possible than others. Jacob Bernoulli had explained this in his *Ars Conjectandi* ([12], page 322 of Edith Sylla's translation):

> I assume that all cases are equally possible, or can happen with equal ease. Otherwise a correction must be made. For any case that happens more easily than others as many or more cases must be counted as it more easily happens. For example, in place of a case three times as easy I count three cases each of which may happen as easily as the rest.

When we are considering a continuous quantity, this selective subdivision can be accomplished by a transformation – what we now call a reparametrization.

14. The passage quoted is in Chapter 4, §17, "The Bases of Laplace's Theory lie in an Experience as to Ignorance." Laplace's law of succession was integral to Pearson's philosophy of science. He distinguished between perceptions (sense impressions) and conceptions (theories), and he saw the law of succession in our construction of theories from repeated perceptions. This view echoes Laplace's associationist psychology [47, 220].

    An avid student of German culture [176], Pearson called his philosophy "a sober idealism" in the preface to the second edition of *The Grammar of Science*. Harold Jeffreys, the best known proponent of inverse probability in England in the 1930s, stated in the preface to the third edition of his *Theory of Probability* [122] that *The Grammar of Science* was his primary inspiration. Jerzy Neyman, perhaps the most influential proponent of Bernoullian statistics in the 20th century, stated in 1957 [165] that he had learned from the *The Grammar of Science* "that scientific theories are no more than models of natural phenomena, frequently inadequate models."

15. Although Fisher directly criticized Bayes's 1763 article in his 1922 article, he eventually convinced himself that Bayes shared his own understanding of probability. He then contended that the Laplace had introduced the aspect of inverse probability to which he objected: the use of prior probabilities unsupported by frequency evidence [4].

16. This story has been recounted by a number of authors; see for example [86]. A discussion led by Leonard J. Savage at Birbeck College in 1959 [6] gives some insight into how the new subjectivism looked to mathematical statisticians at the time.

    The introduction of the terms *Bayesian* and *Bayesianism* is discussed in Appendix I. The use of *Bayesian* as an adjective in English can be found as early as 1948, but its use as a noun does not appear until the 1960s. Nor does *Bayesianism*. Before then, no one called themselves Bayesian.

    The name *Bayesian* has now also been adopted by authors who continue to advocate Laplacean inverse probability, with various ways of justifying the objectivity of the prior probabilities. These include Edwin Jaynes [121], Roger Rosenkrantz [182], James Berger [11], and Jon Williamson [240]. The name *objective Bayesian*, perhaps first used by Rosenkrantz, is now often used for this group.

17. For discussions of other ways Ramsey, de Finetti, and others have tried to understand unknown probabilities in subjective terms, see [99, 143, 152].

18. The point is more or less implicit in some of Laplace's examples, but Laplace did not state it explicitly. As we have seen, he was more concerned with the fact that even the form of the probability laws is unknown, and he seldom bothered with explicit prior probabilities.

    Cournot, whom we may call the first Bernoullian because of the clarity of his early critique of inverse probability, addressed the question as follows in 1843, considering the case where balls are drawn from an urn with unknown numbers of black and white balls:

    > When the numbers [drawn] ... are very large ... the result obtained from Bayes's rule will no longer differ noticeably from the result obtained from Bernoulli's theorem. Well they should, because the validity of Bernoulli's theorem is independent of any hypothesis concerning the initial choice of an urn. In this case it is not (as many authors have apparently thought) Bernoulli's rule that becomes exact by approaching Bayes's rule; it is Bayes's rule that becomes exact,

> or acquires an objective value that it did not have, by becoming the same as Bernoulli's rule.

Here are Cournot's words in the original French ([32], Section 95):

> Quand les nombres . . . sont très grands . . . le résultat trouvé par la règle de Bayes ne diffère plus sensiblement de celui que donnerait le théorème de Bernoulli. Il faut bien qu'il en soit ainsi, puisque la vérité du théorème de Bernoulli est indépendante de toute hypothèse sur le triage préalable de l'urne. Ce n'est point dans ce cas (comme beaucoup d'auteurs ont paru se le figurer) la règle de Bernoulli qui devient exacte en se rapprochant de la règle de Bayes; c'est la règle de Bayes qui devient exacte, ou qui acquiert une valeur objective qu'elle n'avait pas, en se confondant avec la règle de Bernoulli.

Cournot went on to note that if the urn from which the balls are drawn is itself chosen at random, and the chances of getting urns with various proportions of black and white balls is unknown, then Bayes's rule (which assumes a uniform prior distribution) will give the wrong answer, but that it the difference will the negligible when the number of balls drawn in sufficiently large.

Edgeworth took up the issue in 1884 ([76], pages 228–229). Citing Cournot and silently repurposing his argument (taking the distribution over urns with different proportions of black and white balls to be known and subjective rather than unknown and objective), Edgeworth wrote:

> There is not required a precise *à priori* knowledge. . . Almost any *à priori* knowledge, as Cournot has well shown, is sufficient to deduce an overwhelmingly large, though not of course a numerically-measured, probability.

Edgeworth elaborated and repeated his own version of the argument several times; see [80] and the references he gives on page 83 of that article. Arthur Lyon Bowley picked the argument up from Edgeworth and explained it in his widely used textbook, *The Elements of Statistics* [19], beginning with its fourth edition in 1920 (page 414).

Some authors now lump Edgeworth's conclusion, that the subjective prior will not matter if it is smooth when there are many observations, together with Laplace's 1810 conclusion that the posterior resulting from his inverse principle (where the a uniform prior is implicit) will be approximately normal and centered on the true value of the parameter), under the name "Bernstein–von Mises theorem". Several modern versions of this theorem have been developed, but they all require that the number of observations must be many times the number of parameters, and either the parameter space Θ must be finite or stringent conditions must be imposed on the class of smooth prior distributions considered [98].

There is little justification for the name "Bernstein–von Mises theorem". Apparently the name was first used in print in 1956 by Lucien Le Cam, then a junior faculty member in statistics at Berkeley [136]. To all appearances, however, the name is due to Jerzy Neyman, who explained in 1962 [167] that he had learned the result personally from Sergei Bernstein as a student in 1915 or 1916 and cited a 1919 article by Richard von Mises [228]. In fact, von Mises discusses the result only for the binomial case, where it was already proven by Laplace, and cites Czuber's textbook for the result ([228], page 84 , [43], page 218 of the 3rd edition of, 1914). Le Cam [135] cites the first edition of Bernstein's probability textbook, published in 1917. I have not seen this edition, but the 1927 edition discusses only Laplace's Bernoullian result for the binomial and does not discuss Bayesian asymptotics.

19. The situation is somewhat different in some other disciplines. In economics, interest in mathematical models of economic agents keeps the Bayesian picture in view as the default model of rationality. Philosophers continue to use it as a general model of uncertain reasoning. Perceptual psychologists are now using it in detailed studies of how past experience is combined with inputs from the perceptual system.

20. This name *structural equation* was used by Donald A. G. Fraser and others beginning in the 1960s; see [97]. Some recent authors use instead the name *data generating equation*. I prefer to avoid this name, because it suggests that u and therefore x are fully or partly created by chance, perhaps by the goddess called Tyche by the Greeks and Fortuna by the Romans.

Montmort and De Moivre believed they had conquered this pagan goddess [9], but she seems to have regained her powers in recent decades.

21. In 1958, in a letter to John Tukey in 1958 ([10], page 233), Fisher explained that he used the word probability "in the sense in which it was used by the old masters, Fermat, Pascal, Leibnitz, Bernoulli, Montmort, de Moivre, and Bayes." These old masters, like even earlier authors who calculated odds in games of chance [8] saw multiple meanings in the chances they counted; these chances are subjective, because they define our bets, but also objective, because they play out in what happens. It is not surprising, therefore, that different readers of Fisher see different aspects of his understanding of probability. Dempster argues that he understood probability primarily as degree of rational belief [68], Lehmann contends that he understood it primarily as frequency ([141], and Zabell sees some vacillation ([246], pages 83–86 and pages 371–374 and 381).

    For a late essay by Fisher on the nature of probability, see [96].

22. See Fisher's 22 March 1955 letter to Georges Darmois and his 28 March 1957 letter to E. B. Wilson in [10], pages 79–80 and 239.

23. As von Mises explained, failure of the irregularity axiom would allow an opponent to make money betting against the probability [230].

    Related ideas go back at least to John Venn's *Logic of Chance*, published in 1866 [225]. As Venn pointed out, any individual belongs to many classes, and when we interpret the probability that the individual has a given feature as the frequency of that feature in a class, we must decide on the class. Hans Reichenbach called this the problem of selecting a *reference class* [178]. We do not want to use a given class if we can see that the individual belongs to a smaller class in which the frequency is different.

    The idea can also be expressed in terms of subjective probability: with respect to the feature in question the individual should be *exchangeable* with the other individuals in the class [49].

24. See [95]. There are also comments in this direction in the third (posthumous) edition of *Statistical Methods and Scientific Inference* (1973).

25. See [71] and the web site for the Belief Function and Applications Society, `http://www.bfasociety.org/`. For an accounting of my own work on Dempster-Shafer belief functions in the 1970s and 1980s and its relation to my later work, see [200].

26. Often cited is Charles Stein's 1959 example of the discrepancy between fiducial and Bernoullian estimates of the sum of squares of many normal means [211]. In this example, $x_1, \ldots, x_n$ are normal and independent with unit variances and means $\theta_1, \ldots, \theta_n$. We set $h(\theta) := \theta_1^2 + \cdots \theta_n^2$ and $d^2 := x_1^2 + \ldots + x_n^2$, and we propose to estimate $h(\theta)$ using d^2. Because d^2 has expected value $h(\theta) + n$ and variance $2n + 4h(\theta)$, a Bernoullian analysis gives high probability to a confidence interval for $h(\theta)$ of width of order $\sqrt{n}$ around $d^2 - n$. Fisher's fiducial argument for this model produces a probability distribution for $h(\theta)$ that has mean $d^2 + n$ and variance $2n + 4d^2$, which gives high probability to an interval of width of order $\sqrt{n}$ around $d^2 + n$. A Bayesian analysis using a prior that is flat in a very large region of $\mathbb{R}^n$ that turns out to have $x_1, \ldots, x_n$ well in its interior will give approximately the same results as Fisher's fiducial argument.

27. The name *confidence interval* was introduced by Jerzy Neyman in an effort to explain what he thought Fisher was doing with his fiducial intervals, but the idea goes back to Laplace [140]. Cournot explained how the idea is independent of Bayesian thinking (see [32], Section 107, quoted in Section 3.1 below). Aldrich discusses its use by 20th century authors from whom Fisher drew inspiration [2].

28. For example, Hannig has suggested to me that in Stein's example, described in Footnote 8 above, an appropriate data generating function for the feature $h(\theta) := \theta_1^2 + \cdots + \theta_n^2$ might be based on the inverse of the cumulative distribution function for the non-central chi-squared distribution.

29. A *data-dependent* prior distribution is one chosen after the likelihood function is observed. It is inconsistent with the rationale for Bayesian reasoning to tailor the prior to be consistent with the likelihood, but some statisticians systematically do this, especially if an initially chosen prior conflicts strongly with the likelihood. Some authors, such as George E. P. Box [20], have defended an iterative process of Bayesian calculation, model checking, and adjusting the prior.

30. Initially, in his 1930 article, Fisher suggested that fiducial probabilities are probabilities of a different kind. But he soon changed his mind, arguing that they are probabilities like any other, and that they differ from Bayesian posterior probabilities (at least the ones he thought legitimate, those where the prior distribution expresses frequencies in a population from which θ is drawn) only in the argument that produces them.

31. Permit me to deny, without repeating arguments I have made elsewhere (in [197], for example), the claim that a rational person should have already integrated all of his or her evidence and can find the resulting probabilities by examining his or her dispositions to act.

32. As de Finetti wrote in 1937, "the degree of probability attributed by a given individual to a given event" can be defined by "making mathematically precise the trivial and obvious idea" that it "is revealed by the conditions under which he would be disposed to bet on that event" ([49], page 6). In later work, de Finetti argued that it would be more operational to ask individuals to choose between certain loss functions ([56], Chapter 5).

33. This formulation is related to the game-theoretic version of Cournot's principle; see [198] and Sections 3.4 and 5.2 and Appendix II below.

34. Bernoulli considered the difference between the estimated odds $(y/n)/(1-(y/n))$ and the true odds $p/(1-p)$ rather than the difference between the estimated probability y/n and the true probability p. Moreover, his calculation can be improved; considerably smaller values of n than the ones he found will do. The important point, as he emphasized, is that we can find such values; statistical estimation is possible in principle. See Stigler 1986 [215], Chapter 2.

35. This was the first central limit theorem. Again see Chapter 2 of [215] for details.

36. Here is the French original of the translated passage ([32], Section 107):

> La probabilité P a, comme nous l'avons expliqué, une valueur objective; elle mesure effectivemnent la probabilité de l'erreur du jugement que nous portons, en prononçant que la différence $\left|p - \frac{n}{m}\right|$ tombe entre les limites $\pm l$. Lors même que, dans la multitude indéfinie de faits auxquels peuvent s'appliquer les observations statistiques, des raisons inconnues rendraient certains valeurs de p habiles à se produire plus fréquemment que d'autres, le nombre des jugements vrais que nous émettrions, en prononçant, d'après la probabilité P, que la différence $\left|p - \frac{n}{m}\right|$ tombe entre les limites $\pm l$, serait au nombre des jugements erronés sensiblement dan le rapport de P à 1 − P, si d'ailleurs on embrassait une série de jugements assez nombreux pour que les anomalies fortuites aient dû se compenser sensiblement.

37. Neyman introduced the term *confidence interval* in English in 1934 [162]. It is often thought that he developed the definition and introduced the name in reaction to Fisher's fiducial argument, but in 1941 [164] he explained that he had developed and used the concept in Poland around 1930, without knowing about Fisher's work, and that the name *confidence interval* was a translation of the Polish *przedział ufności*.

38. Neyman wrote to de Finetti in French:

> L'expression "coefficient de confiance" employée par moi désigne une *valeur* de la probabilité pour qu'une estimation soit correcte, une valeur choisit arbitrairement d'avance; par conséquent cette expression n'est pas un synonyme du terme "probabilité".

De Finetti quotes him on page 29 of [50].

39. I hasten to repeat that De Moivre had no such notation.

40. For example, if Player I announces the probability 0.05 for A, then Player II is allowed to bet on A at the odds $1:19$. By betting 5 cents on A, he increases the 5 cents to 1 dollar if A happens and loses only the 5 cents if A fails.

41. De Finetti made the point in this way ([56], Section 11.2.2):

> ... the H appearing in $\mathbf{P}(E|H)$ means that this is the probability You attribute to E if 'in addition to your present information ... *it will become known to You that H is true (and nothing else)*'. It would be wrong, therefore, to state, or to think, in a superficial manner, without at least making sure that these explanations are implicit, that $\mathbf{P}(E|H)$ is the probability of E once H is known. In general, by the time we learn that H has occurred, we will already have learnt of other circumstances which might also influence our judgement. In any case, the evidence which establishes that H has occurred will itself contain, explicitly or implicitly, a wealth of further detail, which will modify our final state of information, and, most likely, our probabilistic judgement.

42. Price reported (page 371) that Bayes thought that he could solve the problem ...

> ... provided some rule could be found according to which we ought to estimate the chance that the probability for the happening of an event perfectly unknown, should lie between any two named degrees of probability, antecedently to any experiments made about it; and that it appeared to him that the rule must be to suppose the chance the same that it should lie between any two equidifferent degrees; which, if it were allowed, all the rest might be easily calculated in the common method of proceeding in the doctrine of chances. Accordingly, I find among his papers a very ingenious solution of this problem in this way. But he afterwards considered, that the *postulate* on which he had argued might not perhaps be looked upon by all as reasonable ...

43. My translation of a passage in Section 89. See [202] for additional translations from Cournot.

44. The structural equation G can be written in the form $G(u,p) =$ happen if $u \leq p$ and $G(u,p) =$ fail if $u > p$, where u is uniformly distributed on $[0,1]$.

45. Here Dempster's and my views have diverged. In 1968, Dempster observed that "the connection [between probability and betting] is so close that it is almost of the nature of a tautology to speak of one or of the other" ([64], page 244). He now emphasizes a logical conception of probability not based on betting, in the tradition of De Morgan, Boole, and Jevons [69, 226], whereas I now take a betting version of Cournot's principle (see Section 5) as basic to the meaning of probability, and this is more Bernoullian than logical. And whereas Dempster now disavows phrases such as "continuing to believe", I see continuing to trust (that a gambling strategy using given odds will not multiply one's capital by a large factor) as the best way to express the judgement of independence or irrelevance needed for Dempster's rule.

46. The notion of "conditioning" or "conditionalizing" a probability distribution is fairly recent. To the best of my knowledge, the word *conditionalize* was first used in connection with probability by William K. Estes and Patrick Suppes in 1957 [84]. I prefer *condition* to *conditionalize*, and in my 1976 book [190], I used the name *Bayes's rule of conditioning* for the rule corresponding to Bayes's fifth proposition, the rule that tells us to change the our probability for A from $P(A)$ to $P(A\&B)/P(B)$ when we learn B. This seems to have been the first time this name was used in print. The notion of *conditioning* or *conditionalizing* a probability distribution plays little or no role in mathematical probability theory, because the notion of a stochastic process includes a protocol or filtration [195] for how probability changes with time or information, and this excludes the idea that new information might be an arbitrary subset of the probability space. Both Bayes's rule of conditioning (and Dempster's rule of

combination) take us outside the mathematical theory and require a fiducial judgement.

47. And perhaps even farther back. See Chapter 9 of the second book of Laplace's *Théorie analytique* [132].

48. [175], page 478, my translation. The original French: "Les choses de toute nature sont soumise à une loi universelle qu'on peut appeler *la loi des grandes nombres*. Elle consiste en ce que, si l'on observe des nombres très considérables d'événements d'un même nature, dépendants de causes qui varient irrégulièrement, tantôt dans un sens, tantôt dans l'autre, sans que leur variation soit progressive dans aucun sens déterminé, on trouvera, entre ces nombres, des rapports à peu près constants."

49. See Stephen Stigler's summary on page 182–186 of [215]. Stigler regards I. J. Bienaymé as Poisson's most effective critic. See also [27, 119].

50. Lévy focused not on the martingale law of large numbers as stated here but on the corresponding central limit theorem, which approximates the probability in (11) using the normal distribution, thus strengthening the martingale law of large numbers just as just as De Moivre improved on Bernoulli's theorem. See [13, 142, 138].

51. Jerzy Neyman, Abraham Wald, and other immigrants to Britain and the United States were already changing the direction of statistics in the 1930s, but as Sandy Zabell has pointed out, it was only in the 1940s that probability, with the help of newcomers such as William Feller and Mark Kac, became a distinct and powerful branch of mathematics in the United States [247].

52. The role of moral certainty in early probability theory, the introduction of the name *Cournot's principle*, and its use by multiple authors in the 1950s is discussed at greater length in [198, 207, 208]. After first adopting the name *Cournot's principle*, Fréchet later suggested the name *Buffon-Cournot principle*, but this was not followed by other authors.

53. The condition that the criterion for testing a probabilistic theory be chosen in advance was emphasized by Cournot and Borel; see Cournot's discussion of multiple testing translated in [202] and Borel's discussion in his 1914 book, *Le Hasard* [16]. Neyman cited this discussion by Borel as an inspiration for the ideas in his work with E. S. Pearson on hypothesis testing [168, 139]. There is a difference, however, between choosing a rejection region in advance and choosing only a test statistic from which a p-value will be calculated. The game-theoretic framework of [206] provides straightforward ways to correct for the incompleteness of a test statistic as a specification of a test, and the correction is generally comparable to using a Bayesian significance test of Harold Jeffreys's type [202].

54. See page 221 of [51], Sections 5.2.3 and 5.10.9 of [56], and page 163 of [57].

55. My translation of the following passage in French on page 235:

> La définition de la probabilité subjective est basée sur le comportement de celui qui l'évalue: elle consiste dans la mesure des sacrifices qu'il croit convenable d'accepter pour échapper au risque d'un dommage qui surviendrait avec l'événement considéré (taux d'assurance, de pari, etc.). En particulier, dire que la probabilité est petite, signifie que l'on juge le risque comme négligeable, c'est-à-dire, que l'on agit à peu près comme si l'événment était impossible. Si cela n'est plus un principe, c'est qu'il est par définition un synonyme, une tautologie, une banalité.

56. In his later work, de Finetti systematically used the Italian *previsione*, which can be translated into French as *prévision* and into English as *forecast*, for the probabilistic concept of expected value. A probability, being the expected value of a zero-one variable, is also a *previsione*. De Finetti explains his distinction between *previsione* (forecast) and *predizione* (prediction) in [56], Sections 3.1.2 and 5.2.3. De Finetti's *previsione* was rendered as *prevision* in the English

translation of [56], and this use in English has subsequently been adopted by several other authors, including Walley in his work on imprecise probabilities [237].

57. The 19th century uses of *Laplacean* and *Laplacian* in English that I have found are in physics rather than in probability. Arne Fisher, in his 1915 book on probability [88], calls the normal probability curve *Laplacean*. In the preface to the second edition, in 1922, he refers to Laplace's Bernoullian work using the phrases "Laplacean methods" and "Laplacean doctrine of frequency curves".

58. The French phrase *Méthode inverse des probabilités* appeared much earlier in unpublished teaching notes by Fourier; see [38].

59. In one preface, he wrote "Bayesian probabilities *a posteriori*" (page 1.2b), in another "Bayesian probability *a posteriori*" (page 22.527a).

60. *Bernoullian* has also been used in reference to various other contributions by the Bernoullis. In probability theory, it has been used to refer to Daniel Bernoulli's theory of utility and to various aspects of Jacob Bernoulli's problem of estimating a probability from repeated trials.

References

For GTP Working Papers, see `http://probabilityandfinance.com`.

[1] John Aldrich. R. A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3):162–176, 1997. 39

[2] John Aldrich. Fisher's "inverse probability" of 1930. *International Statistical Review*, 68(2):155–172, 2000. 10, 43

[3] John Aldrich. Fisher and regression. *Statistical Science*, 20(4):401–417, 2005. 39

[4] John Aldrich. R. A. Fisher on Bayes and Bayes' theorem. *Bayesian Analysis*, 3(1):161–170, 2008. 41

[5] Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, editors. *Introduction to Imprecise Probabilities*. Wiley, Chichester, 2014. 23, 62

[6] George G Barnard and David R. Cox, editors. *The Foundations of Statistical Inference: A Discussion*. Methuen, London, 1962. 41

[7] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. 5

[8] David R. Bellhouse. *De vetula*: A medieval manuscript containing probability calculations. *International Statistical Review*, 68(2):123–136, 2000. 38, 43

[9] David R. Bellhouse. Banishing Fortuna: Montmort and De Moivre. *Journal of the History of Ideas*, 69(4):559–581, 2008. 43

[10] J. H. Bennett, editor. *Statistical inference: Selected correspondence of R. A. Fisher*. Clarendon, Oxford, 1990. 39, 43

[11] James Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006. 41

[12] Jacob Bernoulli. *Ars Conjectandi*. Thurnisius, Basel, 1713. This pathbreaking work appeared eight years after Bernoulli's death. Bound together with it were two other works by Bernoulli, a treatise in Latin on infinite series, and a study in French of odds in court tennis (*Lettre à un Amy, sur les Parties du Jeu de Paume*). An English translation, by Edith Sylla, was published by Johns Hopkins University Press in 2006 under the title *The Art of Conjecturing*. 41

[13] Serge Bernstein. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Mathematische Annalen*, 97:1–59, 1927. 46

[14] Friedrich Wilhelm Bessel. *Fundamenta astronomiae pro anno MDCCLV*. Regiomonti, 1818. 6

[15] Laurent Bienvenu, Glenn Shafer, and Alexander Shen. On the history of martingales in the study of randomness. *Electronic Journal for History of Probability and Statistics*, 5(1), 2009. 28

[16] Émile Borel. *Le Hasard*. Félix Alcan, Paris, 1914. 46

[17] Émile Borel. *Valeur pratique et philosophie des probabilités*. Gauthier-Villars, Paris, 1939. 28

[18] Émile Borel. *Probabilité et certitude*. Presses Universitaires de France, Paris, 1950. An English translation, *Probability and Certainty*, was published in 1963 by Walker, New York. 28

[19] Arthur Lyon Bowley. *Elements of Statistics*. King, Westminster, 1901. Later editions appeared in 1902, 1907, 1920, 1925, and 1937. 20, 42

[20] George E. P. Box. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A*, 143(4):383–430, 1980. 27, 36, 44

[21] George E. P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley, New York, 1973. 36

[22] Jean Bricmont. *Making Sense of Quantum Mechanics*. Springer, 2016. 37

[23] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950. 28

[24] Bernard Bru. Poisson, le calcul des probabilités, and l'instruction public. In Piere Costabel, Pierre Dugac, and Michel Métiver, editors, *Siméon-Denis Poisson et la science de son temps*, pages 51–94. École Polytechnique, Palaiseau, 1981. English translation in [26]. 24, 48

[25] Bernard Bru. Souvenirs de Bologne. *Journal de la Société Française de Statistique*, 144(1–2):134–226, 2003. 40

[26] Bernard Bru. Poisson, the probability calculus, and public education. *Electronic Journal for History of Probability and Statistics*, 1(2), November 2005. Translation of [24]. 48

[27] Bernard Bru, Marie-France Bru, and Olivier Bienaymé. La statistique critiquée par le calcul des probabilités : deux manuscrits inédits d'Irenée Jules Bienaymé. *Revue d'Histoire des Mathématiques*, 3(2):137–239, 1997. 46

[28] Marie-France Bru and Bernard Bru. *Le jeu de l'infini et du hasard*. Presses Universitaires de Besançon, to appear. 31, 39, 40

[29] Robert J. Buehler and A. P. Fedderson. Note on a conditional property of Student's t. *Annals of Mathematical Statistics*, 34:1098–1100, 1963. 12

[30] Pafnutii Lvovich Chebyshev. Démonstration élementaire d'une proposition générale de la théorie des probabilités. *Journal für die reine und angewandte Mathematik*, 33:259–267, 1846. 24

[31] Marquis de Condorcet. *Arithmétique politique: textes rares ou inédits (1767-1789). Edition critique commentée par Bernard Bru et Pierre Crépel*. Presses universitaires de France, Paris, 1994. 40

[32] Antoine Augustin Cournot. *Exposition de la théorie des chances et des probabilités*. Hachette, Paris, 1843. Reprinted in 1984 as Volume I (Bernard Bru, editor) of [33]. 6, 9, 42, 43, 44

[33] Antoine Augustin Cournot. *Œuvres complètes*. Vrin, Paris, 1973–2010. The volumes are numbered I through XI, but VI and XI are double volumes. 49

[34] David R. Cox. *Principles of Statistical Inference*. Cambridge University Press, 2006. 37

[35] David R. Cox and David V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974. Second edition 1979. 36

[36] Harald Cramér. *Mathematical Methods in Statistics*. Princeton University Press, Princeton, NJ, 1946. 34

[37] Pierre Crépel. Condorcet, la théorie des probabilités et les calculs financiers. In Roshdi Rashed, editor, *Sciences à l'époque de la Révolution française*, pages 267–325. Blanchard, Paris, 1988. 40

[38] Pierre Crépel. De Condorcet à Arago : l'enseignement des probabilités en France de 1786 à 1830. *Bulletin de la SABIX*, 4:29–55, 1989. 47

[39] William Morgan Crofton. Probability. *Encyclopædia Brittanica, Ninth Edition*, XIX:768–788, 1885. 22

[40] Emanuel Czuber. *Theorie der Beobachtungsfehler*. Teubner, Leipzig, 1891. 40

[41] Emanuel Czuber. *Die Entwickung der Wahrscheinlichkeitstheorie und ihrere Anwendungen*. Teubner, Leipzig, 1899. This book was issued as part two of volume 7 of the *Jahresbericht der Deutschen Mathematiker-Vereinigung*. 40

[42] Emanuel Czuber. Wahrscheinlichkeitsrechnung. In *Encyklopädie der mathematischen Wissenschaften, Band I, Teil 2*, pages 733–767. Teubner, Leipzig, 1900. 30

[43] Emanuel Czuber. *Wahrscheinlichkeitsrechnung und ihre Anwendung auf Fehlerausgleichung, Statistik und Lebensversicherung*. Teubner, Leipzig, 1903. The preface is dated November 1902. Later editions were in two volumes. The two volumes for the second edition appeared in 1908 and 1910, respectively. The third edition of the first volume appeared in 1914. 30, 40, 42

[44] Emanuel Czuber. *Die philosophischen Grundlagen der Wahrscheinlichkeitsrechnung*. Teubner, Leipzig and Berlin, 1923. 31

[45] Andrew W. Dale. *A History of Inverse Probability from Thomas Bayes to Karl Pearson*. Springer, New York, second edition, 1999. 40

[46] Donald A. Darling. Review of *Stochastic analysis of stationary time series*, by Ulf Grenander and Murray Rosenblatt. *Bulletin of the American Mathematical Society*, 64(2):70–71, 1958. 34

[47] Lorraine Daston. *Classical Probability in the Enlightenment*. Princeton University Press, Princeton, NJ, 1988. 41

[48] A. P. Dawid, Steven de Rooij, Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Insuring against loss of evidence in game-theoretic probability, 2010. GTP Working Paper 34. A version appeared in *Statistics and Probability Letters* 81:157–162, 2011. 38

[49] Bruno de Finetti. La prévision, ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937. An English translation by Henry E. Kyburg, Jr., is included in both editions of [130]. 43, 44

[50] Bruno de Finetti. *Compte rendu critique du colloque de Genève sur la théorie des probabilités*. Number 766 in *Actualités Scientifiques et Industrielles*. Hermann, Paris, 1939. This is the eighth fascicle of [238]. 45

[51] Bruno de Finetti. Recent suggestions for the reconciliation of theories of probability. In Jerzy Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 217–225. University of California Press, Berkeley, 1951. 46

[52] Bruno de Finetti. Rôle de la théorie des jeux dans l'économie et rôle des probabilités personelles dans la théorie des jeux. In *Colloque international sur les fondements et applications de la théorie du risque*, volume 40 of *Colloques internationaux*, pages 49–63. C.N.R.S., Paris, 1953. 7

[53] Bruno de Finetti. Media di decisioni e media di opinioni. *Bulletin de l'Institut International de Statistique*, 34:144–157, 1954. 28th session, Part 2. 32

[54] Bruno de Finetti. La notion de "horizon bayesien". In Centre belge de recherches mathématiques, editor, *Colloque sur l'analyse statistique: tenu à Bruxelles le 15, 16 et 17 décembre, 1954*, pages 57–71. Masson, Liège, 1955. 32

[55] Bruno de Finetti. Notes de M. B. de Finetti sur le "Rapport général", 1955. Pages 232–241 of *Les mathématiques et le concret*, by Maurice Fréchet, Presses Universitaires de France. 28

[56] Bruno de Finetti. *Teoria Delle Probabilità*. Einaudi, Turin, 1970. An English translation, by Antonio Machi and Adrian Smith, was published as *Theory of Probability* by Wiley (London, England) in two volumes in 1974 and 1975. 44, 45, 46, 47

[57] Bruno de Finetti. *Probability, Induction and Statistics: The Art of Guessing*. Wiley, London, 1972. Ten articles in English, all previously published in English, French, or Italian in the period from 1949 to 1967. 46

[58] Arthur P. Dempster. New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, 37:355–374, 1963. 12

[59] Arthur P. Dempster. Further examples of inconsistencies in the fiducial argument. *Annals of Mathematical Statistics*, 34(3):884–891, 1966. 12, 23, 32

[60] Arthur P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967. 12

[61] Arthur P. Dempster. Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrika*, 38:512–528, 1967. 12

[62] Arthur P. Dempster. Crosscurrents in statistics; Review of *The Selected Papers*, by E. S. Pearson, *Joint Statistical Papers*, by Jerzy Neyman and E. S. Pearson, and *A Selection of Early Statistical Papers*, by J. Neyman. *Science*, 160:661–663, 1968. 32

[63] Arthur P. Dempster. A generalization of Bayesian inference (with discussion). *Journal of the Royal Statistical Society B*, 30:205–247, 1968. 12

[64] Arthur P. Dempster. The theory of statistical inference: A critical analysis. Chapter 2. Probability. Research Report S-3, Department of Statistics, Harvard University, September 1968. 45

[65] Arthur P. Dempster. Upper and lower probabilities generated by a random closed interval. *Annals of Mathematical Statistics*, 39:957–966, 1968. 12

[66] Arthur P. Dempster. Upper and lower probability inferences for families of hypotheses with monotone density ratio. *Annals of Mathematical Statistics*, 40:953–969, 1969. 12

[67] Arthur P. Dempster. Model searching and estimation in the logic of inference. In V. P. Godambe and D. A. Sprott, editors, *Foundations of Statistical Inference*, pages 56–81. Holt, Rinehart and Winston of Canada, Toronto, 1971. 36

[68] Arthur P. Dempster. Bayes, Fisher, and belief functions. In Seymour Geisser, James S. Hodges, S. James Press, and Arnold Zellner, editors, *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George Barnard*. North-Holland, 1990. 23, 43

[69] Arthur P. Dempster. Logicist statistics. I. Models and modeling. *Statistical Science*, 13(3):248–276, 1998. 45

[70] Arthur P. Dempster. Statistical inference from a Dempster-Shafer perspective. In Xihong Lin, Christian Genest, David L. Banks, Geert Molenberghs, David W. Scott, and Jane-Ling Wang, editors, *Past, Present, and Future of Statistical Science*. Chapman and Hall/CRC, 2014. 13

[71] Thierry Denœux. 40 years of Dempster-Shafer theory. *International Journal of Approximate Reasoning*, 79:1–6, 2016. 43

[72] Thierry Denœux, Shoumei Li, and Songsak Sriboonchitta. Frequency-calibrated belief functions: A review, 2016. Submitted to the *International Journal of Approximate Reasoning*. 17

[73] Emile Dormoy. Théorie mathématique des jeux de bourse. *Compte rendu des travaux de l'Association française pour l'avancement des sciences*, 16(2):214–215, 1888. 39

[74] Antony Eagle, editor. *Philosophy of Probability: Contemporary Readings*. Routledge, Oxford, 2011. 54

[75] Francis Y. Edgeworth. *À priori* probabilities. *Philosophical Magazine*, 18(112):204–210, 1884. 3

[76] Francis Y. Edgeworth. The philosophy of chance. *Mind*, 9:223–235, 1884. 6, 7, 42

[77] Francis Y. Edgeworth. On the probable errors of frequency-constants. *Journal of the Royal Statistical Society*, 71:381–397, 499–512, 651–678, 1908. 39

[78] Francis Y. Edgeworth. Addendum on probable errors on frequency-constants. *Journal of the Royal Statistical Society*, 72:81–90, 1909. 39

[79] Francis Y. Edgeworth. Mathematical representation of statistics: A reply. *Journal of the Royal Statistical Society*, 81(2):322–333, 1918. 32

[80] Francis Y. Edgeworth. Molecular statistics. *Journal of the Royal Statistical Society*, 84(1):71–89, 1921. 42

[81] A. W. F. Edwards. What did Fisher mean by "inverse probability" in 1912–1922? *Statistical Science*, 12:177–184, 1997. 31

[82] Ward Edwards, Harold Lindman, and Leonard J. Savage. Bayesian statistical inference for psychologists. *Psychological Review*, 70:193–242, 1963. 9, 33

[83] Bradley Efron. Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80:3–26, 1993. 17

[84] William K. Estes and Patrick Suppes. Foundations of statistical learning theory, I. The linear model for simple learning. Technical Report 16, Behavioral Sciences Division, Applied Mathematics and Statistics Laboratory, Stanford University, November 1957. 45

[85] Richard William Farebrother. *Fitting Linear Relationships: A History of the Calculus of Observations, 1750–1900*. Springer, New York, 1998. 39, 40

[86] Stephen E. Fienberg. When did Bayesian inference become Bayesian? *Bayesian Analysis*, 1(1):1–40, 2006. 31, 41

[87] Hans Fischer. *A History of the Central Limit Theorem: From Classical to Modern Probabilty Theory*. Springer, New York, 2011. 39

[88] Arne Fisher. *The Mathematical Theory of Probabilities and Its Application to Frequency Curves and Statistical Methods*. Macmillan, New York, 1915. Second edition 1922. 39, 40, 47

[89] Ronald A. Fisher. The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85:597–612, 1922. 39

[90] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London (A)*, 222:309–368, 1922. 2

[91] Ronald A. Fisher. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22(5):700–725, 1925. 39

[92] Ronald A. Fisher. The logic of inductive inference (with discussion). *Journal of the Royal Statistical Society*, 98(1):39–82, 1935. 39

[93] Ronald A. Fisher. *Contributions to Mathematical Statistics*. Wiley, New York, 1950. 31

[94] Ronald A. Fisher. *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh, 1956. Second edition in 1959, posthumous third edition in 1973. 12

[95] Ronald A. Fisher. The underworld of probability. *Sankhya*, 18:201–210, 1957. 43

[96] Ronald A. Fisher. The nature of probability. *The Centennial Review of Arts & Science*, 2:261–274, 1958. 27, 43

[97] Donald Alexander Stuart Fraser. *The Structure of Inference*. Wiley, New York, 1968. 42

[98] David A. Freedman. Wald lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Annals of Statistics*, 27(4):1119–1141, 1999. 42

[99] Maria Carla Galavotti. New prospects for pragmatism: Ramsey's constructivism. In M. C. Galavotti et al., editor, *New Directions in the Philosophy of Science, The Philosophy of Science in a European Perspective*, pages 645–656. Springer International Publishing Switzerland, 2014. 41

[100] Thomas Galloway. *A Treatise on Probability: Forming the article under that head in the seventh edition of the Encyclopædia Britannica*. Adam and Charles Black, Edinburgh, 1839. 6

[101] Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66:8–38, 2013. 27

[102] Meyer A. Girshick. Review of *Facts from Figures* by M. J. Moroney. *Journal of the American Statistical Association*, 48(263):645–647, 1953. 34

[103] Sheldon Goldstein. Boltzmann's approach to statistical mechanics. In Jean Bricmont, Detlef Dürr, Maria C. Galavotti, Giancarlo Ghirardi, Francesco Petruccione, and Nino Zanghi, editors, *Chance in Physics: Foundations and Perspectives*, Lecture Notes in Physics 574, pages 38–54. Springer-Verlag, 2001. 37

[104] Irving J. Good. *Probability and the Weighing of Evidence*. Hafner, 1950. 33

[105] Irving J. Good. Review of *Theory of Games and Statistical Decisions*, by D. Blackwell and M. A. Girschick. *Journal of the American Statistical Association*, 51(274):388–390, 1956. 33

[106] Irving J. Good. Significance tests in parallel and in series. *Journal of the American Statistical Association*, 53(284):799–813, 1958. 31, 33

[107] Irving J. Good. *Good Thinking*. University of Minnesota Press, 1983. 31

[108] Prakash Gorroochurn. *Classic Topics on the History of Modern Mathematical Statistics from Laplace to More Recent Times*. Wiley, New York, 2016. 6, 39, 40

[109] Peter D. Grünwald. Safe probability. *Journal of Statistical Planning and Inference*, to appear. `http://arxiv.org/abs/1604.01785`. 23

[110] Ian Hacking. *The Taming of Chance*. Cambridge University Press, New York, 1990. 33

[111] Alan Hájek. Fifteen arguments against hypothetical frequentism. In Eagle [74], pages 410–432. 36

[112] Alan Hájek. "Mises redux"-redux: Fifteen arguments against finite frequentism. In Eagle [74], pages 395–409. 36

[113] Jaroslav Hájek. On basic concepts of statistics. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pages 139–162, Berkeley, California, 1967. University of California Press. 35

[114] Anders Hald. *A History of Mathematical Statistics from 1750 to 1930*. Wiley, New York, 1998. 6, 39

[115] Anders Hald. On the history of maximum likelihood in relation to inverse probability and least squares. *Statistical Science*, 14(2):214–222, 1999. 39

[116] Anders Hald. *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713 to 1935*. Springer, New York, 2007. 20, 39, 40

[117] 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: Mathematical, Physical and Engineering Sciences*, 373(2039), 2015. 39

[118] Jan Hannig, Hari Iyer, Randy C. S. Lai, and Thomas C. M. Lee. Generalized fiducial inference: A review. *Journal of the American Statistical Association*, 111:1346–1361, 2016. 17

[119] Christopher C. Heyde and Eugene Seneta. *I. J. Bienaymé: Statistical theory anticipated*. Springer, New York, 1977. 46

[120] Joseph L. Hodges and Erich L. Lehmann. Estimates of location based on rank tests. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 598–611, Berkeley, California, 1961. University of California Press. 34

[121] Edwin T. Jaynes. *Probability theory: The logic of science*. Cambridge, 2003. 41

[122] Harold Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, 1939. Second edition 1948, third 1961. 9, 41

[123] Geoffrey H. Jowett. Review of *An Introduction to Stochastic Processes* by M. S. Bartlett. *Journal of the Royal Statistical Society C*, 5(1):70, 1956. 34

[124] Geoffrey H. Jowett. Statistical analysis using local properties of smoothly heteromorphic stochastic series. *Biometrika*, 44(3/4):454–463, 1957. 34

[125] Maurice G. Kendall. On the reconciliation of theories of probability. *Biometrika*, 36:101–116, 1949. 35

[126] A(leksandr) Khinchine. Über einen Satz der Wahrscheinlichkeitsrechnung. *Fundamenta Mathematicae*, VI:9–20, 1924. The date "December 1922" is given at the end of the article. 27

[127] Judy Klein. *Statistical Visions in Time: A History of Time Series Analysis, 1662–1938*. Cambridge University Press, Cambridge, 1997. 39

[128] Andrei N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933. An English translation by Nathan Morrison appeared under the title *Foundations of the Theory of Probability* (Chelsea, New York) in 1950, with a second edition in 1956. 35

[129] Andrei N. Kolmogorov. Определение центра рассеивания и меры точности по ограниченному числу наблюдений (On the statistical estimation of the parameters of the Gaussian distribution). Известия Академии Наук СССР, Серия математическая *(Bulletin of the Academy of Sciences of the USSR, Mathematical Series)*, 6(1/2):3–32, 1942. 12, 34

[130] Henry E. Kyburg, Jr. and Howard E. Smokler, editors. *Studies in Subjective Probability*. Wiley, New York, 1964. A selection of readings, ranging chronologically from John Venn in 1888 to Leonard J. Savage in 1961. A second edition, with a slightly different selection, was published by Krieger, New York, in 1980. 50

[131] Pierre Simon Laplace. Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Académie royale des sciences de Paris*, 6:621–656, 1774. Reprinted in Volume 8 of Laplace's *Oeuvres complètes*, pages 27–65, translated into English with commentary by Stigler in [216, 134]. 4, 40, 56

[132] Pierre Simon Laplace. *Théorie analytique des probabilités*. Courcier, Paris, first edition, 1812. This monumental work had later editions in 1814 and 1820. The third edition was reprinted in Volume 7 of Laplace's *Œuvres complètes*. 46

[133] Pierre Simon Laplace. *Essai philosophique sur les probabilités*. Courcier, Paris, first edition, 1814. The fifth and definitive edition appeared in 1825 and was reprinted in 1986 (Christian Bourgois, Paris) with a commentary by Bernard Bru. Multiple English translations have appeared. 6, 40

[134] Pierre Simon Laplace. Memoir on the probability of the causes of events. *Statistical Science*, 1(3):364–378, 1986. Translation of [131] by Stigler. 55

[135] Lucien Le Cam. On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics*, 1:277–330, 1953. 42

[136] Lucien Le Cam. On the asymptotic theory of estimation and testing hypotheses. In Jerzy Neyman, editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 129–156. University of California Press, Berkeley, 1956. 42

[137] Lucien Le Cam. Sufficiency and approximate sufficiency. *Annals of Mathematical Statistics*, 35(4):1419–1455, 1964. 35

[138] Lucien Le Cam. The central limit theorem around 1935 (with comments by Hale F. Trotter, Joseph L. Doob, and David Pollard). *Statistical Science*, 1:78–96, 1986. 46

[139] Eric L. Lehmann. The Bertand-Pearson debate and the origins of the Neyman-Pearson theory. In J. K. Ghosh, S. K. Mitra, K. R. Parthasarathy, and B. L. S. Prakasa Rao, editors, *Statistics and Probability: A Raghu Raj Bahadur Festschrift*, pages 371–380. Wiley Eastern, 1993. Reprinted on pages 965–974 of *Selected Works of E. L. Lehman*, edited by J. Rojo, Springer, 2012. 46

[140] Erich L. Lehmann. Some early instances of confidence statements. Technical report, Statistical Laboratory, University of California, Berkeley, September 1958. 31, 43

[141] Erich L. Lehmann. *Fisher, Neyman, and the Creation of Classical Statistics*. Springer, New York, 2011. 35, 43

[142] Paul Lévy. *Théorie de l'addition des variables aléatoires*. Gauthier-Villars, Paris, 1937. Second edition: 1954. 46

[143] David Lewis. A subjectivist's guide to objective chance. In Richard C. Jeffrey, editor, *Studies in Inductive Logic and Probability*, volume II, pages 83–132. University of California Press, 1980. 41

[144] Jean-Baptiste-Joseph Liagre. *Calcul des probabilités et théorie des erreurs avec des applications aux sciences d'observation en général et à la géodésie*. Muquardt, Brussels, 1852. Second edition, 1879, prepared with the assistance of Camille Peny. 39

[145] Dennis V. Lindley. Professor Hogben's 'crisis' – A survey of the foundations of statistics. *Journal of the Royal Statistical Society C, Applied Statistics*, 7(3):186–198, 1958. 31

[146] Dennis V. Lindley. The Bayesian analysis of contingency tables. *Annals of Mathematical Statistics*, 35(4):1622–1643, 1964. 33

[147] Dennis V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge University Press, 1965. Two volumes. 33

[148] Dennis V. Lindley. The choice of variables in multiple regression. *Journal of the Royal Statistical Society B*, 30(1):31–66, 1968. 36

[149] Dennis V. Lindley. The future of statistics: a Bayesian 21st century. *Advances in Applied Probability*, 7:106–115, 1975. 36

[150] Dennis V. Lindley. The 1998 Wald Memorial Lecture: The present position in Bayesian statistics. *Statistical Science*, 5(1):44–65, 1990. 36

[151] Dennis V. Lindley and Adrian F. M. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society B*, 34(1):1–41, 1972. 36

[152] Barry Loewer. David Lewis's Humean theory of objective chance. *Philosophy of Science*, 71:1115–1125, 2004. 41

[153] Andrei Andreevich Markov. *Wahrscheinlichkeitsrechnung*. Teubner, 1912. Translation of second Russian edition. 30

[154] Ryan Martin and Chuanhai Liu. *Inferential models: Reasoning with uncertainty*. CRC Press, Boca Raton, 2016. 17

[155] Ryan Martin, Jianchun Zhang, and Chuanhai Liu. Dempster-Shafer theory and statistical inference with weak beliefs. *Statistical Science*, 25(1):72–87, 2010. 17

[156] Thierry Martin. *Probabilités et critique philosophique selon Cournot*. Vrin, Paris, 1996. 27

[157] Thierry Martin, editor. *Actualité de Cournot*. Vrin, Paris, 2005. 27

[158] Thierry Martin. La réception philosophique de Laplace en France. *Electronic Journal for History of Probability and Statistics*, 8(1), 2012. 31

[159] James Clerk Maxwell. Does the progress of physical science tend to give any advantage to the opinion of necessity (or determinism) over that of the contingency of events and the freedom of the will?, 1873. Pages 362–366 of *The Life of James Clerk Maxwell, with selections from his correspondence and occasional writings and a sketch of his contributions to science*, by Lewis Campbell and William Garnett (MacMillan and Co., London, 1882). 37

[160] Ernest Nagel. The meaning of probability. *Journal of the American Statistical Association*, 31(193):10–30, 1936. 35

[161] Ernest Nagel. *Principles of the Theory of Probability*. University of Chicago Press, 1939. 35

[162] Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934. 44

[163] Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions Royal Society of London, Series A*, 236:333–380, 1937. 11, 35

[164] Jerzy Neyman. Fiducial argument and the theory of confidence intervals. *Biometrika*, 32(2):128–150, 1941. 44

[165] Jerzy Neyman. "Inductive behavior" as a basic concept of philosophy of science. *Review of the International Statistical Institute*, 25(1/3):7–22, 1957. 20, 41

[166] Jerzy Neyman. Indeterminism in science and new demands on statisticians. *Journal of the American Statistical Association*, 55:625–639, 1960. 26

[167] Jerzy Neyman. Two breakthroughs in the theory of statistical decision making. *Review of the International Statistical Institute*, 30(1):11–27, 1962. 42

[168] Jerzy Neyman. Frequentist probability and frequentist statistics. *Synthese*, 36(1):97–131, 1977. 35, 46

[169] Kh. O. Ondar, editor. О теории вероятностей и математической статистике (переписка А. А. Маркова и А. А. Чупрова). Nauk, Moscow, 1977. See [170] for English translation. 19, 58

[170] Kh. O. Ondar, editor. *The Correspondence Between A. A. Markov and A. A. Chuprov on the Theory of Probability and Mathematical Statistics*. Springer, New York, 1981. Translation of [169] by Charles M. and Margaret Stein. Additional letters between Markov are provided in translation by Sheynin in [210], Chapter 8. 58

[171] A. de Forest Palmer. *The Theory of Measurements*. McGraw Hill, New York, 1912. 20

[172] Karl Pearson. *The Grammar of Science*. Scott, London, 1892. A second edition appeared in 1900, a third in 1911. 6

[173] Charles C. Peters and Walter R. Van Voorhis. *Statistical Procedures and their Mathematical Bases*. McGraw-Hill, New York, 1940. 34

[174] Henri Poincaré. *Calcul des probabilités. Leçons professées pendant le deuxième semestre 1893–1894*. Gauthier-Villars, Paris, 1896. Second edition 1912. 20, 39

[175] Siméon-Denis Poisson. Recherches sur la probabilité des jugements, principalement en matière criminelle. *Comptes rendus hebdomadaires des séances de l'Académie des Sciences*, 1:473–494, 1835. Session of 14 December 1835. 46

[176] Theodore M. Porter. *Karl Pearson: The Scientific Life in a Statistical Age*. Princeton University Press, Princeton, NJ, 2004. 41

[177] John W. Pratt. Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society B*, 27(2):169–203, 1965. 9, 33

[178] Hans Reichenbach. Les fondements logiques des probabilités. *Annales de l'Institut Henri Poincaré*, 7:267–348, 1937. 43

[179] Harry V. Roberts. The new business statistics. *The Journal of Business*, 33(1):21–30, 1960. 32

[180] Harry V. Roberts. Review of *Applied Statistical Decision Theory* by Howard Raiffa and Robert Schlaifer. *Journal of the American Statistical Association*, 57(297):199–202, 1962. 32

[181] Paul Romer. The trouble with macroeconomics. `https://paulromer.net/wp-content/uploads/2016/09/WP-Trouble.pdf`, September 14, 2016. To appear in *The American Economist*. 16

[182] Roger D. Rosenkrantz. *Inference, method and decision: towards a Bayesian philosophy of science*. Reidel, Dordrecht, 1977. 41

[183] Leonard J. Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46:55–67, 1951. 31

[184] Leonard J. Savage. The foundations of statistics reconsidered. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 575–586. University of California Press, 1961. 9, 32

[185] Robert Schlaifer. *Probability and Statistics for Business Decisions*. McGraw-Hill, New York, 1959. 34

[186] Tore Schweder and Nils L. Hjort. Confidence and likelihood. *Scandinavian Journal of Statistics*, 29(2):309–332, 2002. 17

[187] Tore Schweder and Nils L. Hjort. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, 2016. 17

[188] Teddy Seidenfeld. R. A. Fisher's fiducial argument and Bayes' theorem. *Statistical Science*, 7(3):358–368, 1992. 12

[189] Stephen Senn. You may believe you are a Bayesian, but you are probably wrong. *Rationality, Markets and Morals*, 2:48–66, 2011. 27

[190] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976. 13, 45

[191] Glenn Shafer. Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences*, 19:309–370, 1978. 13

[192] Glenn Shafer. Constructive probability. *Synthese*, 48:1–60, 1981. 23

[193] Glenn Shafer. Bayes's two arguments for the rule of conditioning. *Annals of Statistics*, 10:1075–1089, 1982. 21

[194] Glenn Shafer. Belief functions and parametric models (with discussion). *Journal of the Royal Statistical Society B*, 44:322–352, 1982. 13

[195] Glenn Shafer. Conditional probability. *International Statistical Review*, 53:261–277, 1985. 45

[196] Glenn Shafer. The combination of evidence. *International Journal of Intelligent Systems*, 1:155–179, 1986. 13

[197] Glenn Shafer. Savage revisited (with discussion). *Statistical Science*, 1:463–501, 1986. 44

[198] Glenn Shafer. From Cournot's principle to market efficiency, March 2006. GTP Working Paper 15. Published as Chapter 4 of: Jean-Philippe Touffut, editor, *Augustin Cournot: Modelling Economics*. Edward Elgar, Cheltenham, UK, 2007. 27, 44, 46

[199] Glenn Shafer. Game-theoretic probability and its uses, especially defensive forecasting, August 2007. GTP Working Paper 22. Published as "Defensive forecasting: How to use similarity to make forecasts that pass statistical tests". Pp. 215-247 of *Preferences and Similarities*, edited by Giacomo Della Riccia, Didier Dubois, Rudolf Kruse, and Hans-Joachim Lenz, CISM Series, Springer, NewYork, 2008. 38

[200] Glenn Shafer. *A Mathematical Theory of Evidence* turns 40. *International Journal of Approximate Reasoning*, 79:7–25, 2016. 43

[201] Glenn Shafer. How speculation can explain the equity premium, November 2016. GTP Working Paper 47. 39

[202] Glenn Shafer. Cournot in English, April 2017. GTP Working Paper 48. 27, 45, 46

[203] Glenn Shafer. Game-theoretic significance testing, April 2017. GTP Working Paper 49. 27

[204] Glenn Shafer, Peter R. Gillett, and Richard B. Scherl. A new understanding of subjective probability and its generalization to lower and upper prevision, October 2002. GTP Working Paper 3. Published in *International Journal of Approximate Reasoning* 31:1–49, 2003. 23

[205] Glenn Shafer and Amos Tversky. Languages and designs for probability judgment. *Cognitive Science*, 9:309–339, 1985. 23

[206] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001. 23, 24, 26, 38, 46

[207] Glenn Shafer and Vladimir Vovk. The sources of Kolmogorov's *Grundbegriffe*. *Statistical Science*, 21:70–98, 2006. 27, 46, 60

[208] Glenn Shafer and Vladimir Vovk. The origins and legacy of Kolmogorov's *Grundbegriffe*, April 2013. GTP Working Paper 4. Abridged version published as "The sources of Kolmogorov's *Grundbegriffe*" [207]. 27, 35, 46

[209] Glenn Shafer, Vladimir Vovk, and Roman Chychyla. How to base probability theory on perfect-information games, 2009. GTP Working Paper 32. First posted in December 2009. Published in BEATCS `http://eatcs.org/index.php/on-line-issues`, 100:115–148, February 2010, pages 115–148). 38

[210] Oscar Sheynin. *Aleksandr A. Chuprov: Life, Work, Correspondence. The making of mathematical statistics*. V&R unipress, Goettingen, 2011. Second revised edition, edited by Heinrich Strecker. The first edition appeared in 1996. 58

[211] Charles Stein. An example of wide discrepancy between fiducial and confidence intervals. *Annals of Mathematical Statistics*, 30:877–880, 1959. 43

[212] Paul J. Steinhardt. The inflation debate: Is the theory at the heart of modern cosmology deeply flawed? *Scientific American*, 304:36–43, April 2011. 37

[213] Stephen M. Stigler. The transition from point to distribution estimation. *Bulletin of the International Statistical Institute*, 46:332–340, 1975. 3

[214] Stephen M. Stigler. Discussion of "On rereading R. A. Fisher", by L. J. Savage. *The Annals of Statistics*, 4(3):498–500, 1976. 39

[215] Stephen M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, MA, 1986. 39, 40, 44, 46

[216] Stephen M. Stigler. Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1(3):359–378, 1986. 40, 55

[217] Stephen M. Stigler. *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA, 1999. 39

[218] Stephen M. Stigler. Fisher in 1921. *Statistical Science*, 20(1):32–49, 2005. 39

[219] Stephen M. Stigler. The epic story of maximum likelihood. *Statistical Science*, 22(4):598–620, 2007. 39

[220] Stephen M. Stigler. Laplace's unpublished manuscript on associationist psychology. *Electronic Journal for History of Probability and Statistics*, 8(1), 2012. 41

[221] Gunnar Taraldsen and Bo H. Lindqvist. Conditional fiducial models. *Statistical Planning and Inference*, to appear. 17

[222] Isaac Todhunter. *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace*. Macmillan, London, 1865. 40

[223] John W. Tukey. Curves as parameters, and touch estimation. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 681–694, Berkeley, California, 1961. University of California Press. 34

[224] Alan M. Turing. The applications of probability to cryptography, c 1941. UK National Archives, HW 25=37. See arXiv:1505.04714 for a version set in Latex. 20

[225] John Venn. *The Logic of Chance: an essay on the foundations and province of the theory of probability, with especial reference to its logical bearings and its application to moral and social science, and to statistics*. Macmillan, London, 1866. 43

[226] Lukas M. Verburgt. The objective and the subjective in mid-nineteenth-century British probability theory. *Historia Mathematica*, 42(4):468–487, 2015. 45

[227] Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939. This differs from Ville's dissertation, which was defended in March 1939, only in that a one-page introduction was replaced by a 17-page introductory chapter. 27

[228] Richard von Mises. Fundamentalsätze der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 4:1–97, 1919. 42

[229] Richard von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5:52–99, 1919. 32, 35

[230] Richard von Mises. *Wahrscheinlichkeitsrechnung, Statistik und Wahrheit*. Verlag von Julius Springer, Vienna, 1928. The second edition appeared in 1936 and the third in 1951. A posthumous fourth edition, edited by his wife Hilda Geiringer, appeared in 1972. English editions, under the title *Probability, Statistics and Truth*, appeared in 1939 and 1957. 43

[231] Vladimir Vovk. Superefficiency from the vantage point of computability. *Statistical Science*, 24(1):73–86, 2009. 39

[232] Vladimir Vovk and Glenn Shafer. Game-theoretic probability. In Augustin et al. [5], pages 114–134. 38

[233] Vladimir Vovk and Glenn Shafer. Basics of a probability-free theory of continuous martingales, July 2016. GTP Working Paper 45. 39

[234] Vladimir Vovk, Akimichi Takemura, and Glenn Shafer. Defensive forecasting, January 2005. GTP Working Paper 8. First posted in September 2004. A version appeared in the AI & Statistics 2005 proceedings. 28

[235] Abraham Wald. *On the principles of statistical inference*. University of Notre Dame, 1942. Four lectures delivered at the University of Notre Dame, Februrary 1941. Printed by Edwards Brothers, Lithoprinters, Ann Arbor. 26

[236] David L. Wallace. The Behrens-Fisher and Fieller-Creasy Problems. In Stephen E. Fienberg and David K. Hinkley, editors, *R. A. Fisher: An Appreciation*, pages 119–147. Springer, 1980. 12

[237] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991. 24, 47

[238] Rolin Wavre. *Colloque consacré à la théorie des probabilités*. Hermann, Paris, 1938–1939. This celebrated colloquium was held in October 1937 at the University of Geneva, as part of a series (*Conférences internationales des sciences mathématiques*) that began in 1933. The colloquium was organized by Rolin Wavre and chaired by Maurice Fréchet. Other participants included Cramér, Doeblin, Feller, de Finetti, Heisenberg, Hopf, Lévy, Neyman, Pòlya, Steinhaus, and Wald, and communications were received from Bernstein, Cantelli, Glivenko, Jordan, Kolmogorov, von Mises, and Slutsky. The proceedings of the colloquium were published by Hermann in eight fascicles of 50 to 100 pages each, in their

series *Actualités Scientifiques et Industrielles*. The first seven fascicles appeared in 1938 as numbers 734 through 740 of this series; the eighth appeared in 1939 as number 766. The second fascicle, entitled *Les fondements du calcul des probabilités*, includes contributions by Feller, Fréchet, von Mises, and Wald. The eighth fascicle consists of de Finetti's summary of the colloquium. 50

[239] Donald Williams. The challenging situation in the philosophy of probability. *Philosophy and Phenomenological Research*, 6(1):67–86, 1945. 35

[240] Jon Williamson. *In Defence of Objective Bayesianism*. Oxford, 2010. 41

[241] Charles P. Winsor. Probability and Listerism. *Human Biology*, 20(3):161–169, 1948. 31

[242] Jacob Wolfowitz. On the theory of runs with some applications to quality control. *The Annals of Mathematical Statistics*, 14(3):280–288, 1943. 34

[243] Min-ge Xie, Regina Y. Liu, C. V. Damaraju, and William H. Olson. Incorporating external information in analyses of clinical trials with binary outcomes. *The Annals of Applied Statistics*, 7(1):342–368, 2013. 16

[244] Min-ge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: A review (with discussion). *International Statistical Review*, 81(1):3–77, 2013. 16, 17

[245] Ronald R. Yager and Liping Liu, editors. *Classic Works of the Dempster-Shafer Theory of Belief Functions*. Springer, Berlin, 2008. 13

[246] Sandy L. Zabell. R. A. Fisher and the fiducial argument. *Statistical Science*, 7(3):369–387, 1992. 12, 43

[247] Sandy L. Zabell. US probability and statistics circa WWII. *Olberwolfach Reports*, 8(4):2945–2948, 2011. 46

[248] Sandy L. Zabell. Commentary on Alan M. Turing: The applications of probability to cryptography. *Cryptologia*, 36(3):191–214, 2012. 20

Acknowledgements

This paper grows out of my comments on Art Dempster's keynote presentation at the Fourth Bayesian, Fiducial, and Frequentist Conference, held at Harvard in May 2017 (`http://statistics.fas.harvard.edu/bff4`; see also `https://sph.umich.edu/biostat/events/bff-conference.html`). The paper has benefited from my conversations and correspondence with numerous participants at this meeting; aside from Dempster, these include Andrew Gelman, Jan Hannig, Ryan Martin, Xiao-Li Meng, Teddy Seidenfeld, Stephen Senn, Steve Stigler, Volodya Vovk, Min-ge Xie, and Sandy Zabell. It has also benefited from my comments by David Bellhouse, Bernard Bru, Pierre Crépel, Thierry Denœux, Maria Carla Galavotti, Shelly Goldstein, Prakash Gorroochurn, Chuanhai Liu, Barry Loewer, Thierry Martin, and Prakash Shenoy.