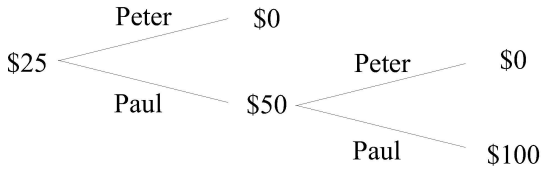
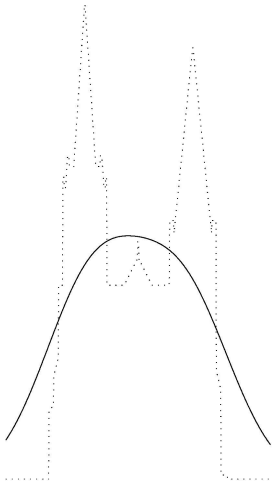
Bayesian, Fiducial, Frequentist

Glenn Shafer
Rutgers University
gshafer@rutgers.edu





The Game-Theoretic Probability and Finance Project

Working Paper #50

First posted April 30, 2017. Last revised August 17, 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 and Bayesians, as well as proponents of fiducial and other probabilistic arguments.

The paper also sketches my understanding of the developments in statistical theory that have led to a renewed interest in fiducial and Dempster-Shafer arguments. It 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`).

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[1] and Bayesians and to reconcile their philosophies with more venturesome ideas provided by R. A. Fisher's fiducial argument and its descendants, including the Dempster-Shafer theory of belief functions.

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 probabilities that are initially subjective or purely theoretical, and to continue to trust them in light of information that we judge materially irrelevant.

- *Poisson's principle*: Even varying probabilities allow probabilistic prediction. The law of large numbers, for example, does not require independent identically distributed trials. The prediction may be about the approximate value of an average, not about the approximate value of the frequency of an event.

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

Each of these principles has venerable historical roots. Each is, in some sense, a truism. But they 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 captures a feature common to Bernoullian and Bayesian statistical practice that brings them both closer to Fisher's fiducial argument than usually thought. Poisson's principle and Cournot's principle can help us see past the widespread but fallacious thesis that Bernoullian statistical theory equates probability with frequency. Instead, it is based on concepts of prediction and testing that many who call themselves Bayesians also use in practice.

In the next section I review recent developments that have led to the current interest in the fiducial argument. In subsequent sections I discuss in turn the three principles that I propose. I have tried to make my account accessible to a relatively wide audience, making the mathematical exposition as elementary as possible and including some historical detail that is well known to many.

[1] Non-Bayesian methods of statistical estimation and testing are now often called *frequentist*. Following Francis Edgeworth, Arthur Dempster, and Ian Hacking, I will instead call them *Bernoullian*, in honor of Jacob Bernoulli. A variety of other names have been used. See Appendix A.

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, which was led by Karl Pearson and included other outstanding contributors such as William S. Gosset and George Udny Yule, had already established international leadership in mathematical statistics. This school's 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" [55, 1, 143]. 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 independent observations to estimate the parameter θ.

As Fisher put it, 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".

This 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 identify the nature of its originality. It captures much of what came before, from Bernoulli's estimation of the probability of an event to Pearson's fitting of frequency curves. Its originality lay perhaps in the extent to which and the way in which it abstracted from practice. For the first time, for example, the constants to be estimated in order to specify a probability law were systemically given a generic name — *parameters* — and were said to describe a hypothetical population "exhaustively in respect of all qualities under discussion".[2]

[2] In 1976 [140], Stephen Stigler noted Fisher's originality in his systematic use of *parameter* and reported finding only a few earlier instances where Fisher's British predecessors had used the word to designate a constant in a probability law. Similar isolated uses of *paramètre* in French include Bravais in 1846 (cited by Edgeworth, as Stigler noted), Liagre in 1852 [92], and Dormoy in 1888 [44]. In neither language was the usage systematic. In Poincaré's textbook [114], for example, *paramètre* is sometimes used for what we now call a random variable (page 98 in the 1896 edition, page 121 in the 1912 edition).

In part, the originality and power of Fisher's framework lay in its narrowness — in 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 two preceding centuries, 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 Karl Pearson, for Pearson and his fellow biometricians were emphasizing random sampling from biological populations. But most statistical work at the beginning of the twentieth century was in fields such as economics, demography, insurance, and meteorology, where time series are central [54, 85]. Even Pearson, Gosset, and Yule contributed to the theory of time series.

For many older statisticians, Fisher's pronouncements concerning the task of theoretical statistics sounded ridiculous. But the study of stochastic processes as a branch of probability theory being in its infancy, the narrowness of Fisher's picture allowed him and his successors to strengthen the probabilistic foundation for statistics, and the success of this mathematical work has kept the picture at the center of thinking about statistics even to this day, sometimes in ways we may not recognize.

It can even be argued that Fisher's narrowing of theoretical statistics opened the way for what we now call Bayesianism. When Fisher came on the scene, Laplace's method of inverse probability was still widely accepted among British philosophers and statisticians, including Karl Pearson. But for most statisticians it was only a method, not an all-encompassing theory of statistics. It was a method for finding the probabilities of possible causes of an event when the probability given the event by each cause is known. There were other uses of probability in statistical work, and even when he wanted to find probabilities for causes, Laplace did not assume that the probability law given by each cause is necessarily known. Indeed, *Laplace's theorem* — the celebrated theorem of 1810 that is now considered an early version of the central limit theorem[3] — was important in his eyes precisely because its allows us to calculate the probabilities of causes when the probability law associated with each cause is not known, provided we have many independent observations.[4] In Fisher's picture, statistical theory is all and only about the probabilities of causes (the possible values of θ being the causes), and the probability law associated with each cause (P_θ) is taken as known, establishing it being the task of the practical statistician, not the theoretical statistician. So the task of theoretical statistics can be completed, it would seem, by specifying prior probabilities for the causes and applying Bayes's rule.

Prior probabilities did not play a large role in Laplace's inverse probability. His principle of inverse probability, probably abstracted from an example worked out by Condorcet[5] and first published in 1774, did not involve prior probabilities;

[3]See Fischer [53], Hald [75, 76, 69].

[4]See [13, 69, 75].

[5]In an unpublished manuscript, Condorcet applied the principle to a problem involving

it simply stated that in light of an event, the probability of each possible cause should be proportional to the probability the event gives the cause. In Fisher's terminology: obtain a probability distribution for the parameter by normalizing the likelihood.[6] After Condorcet and Laplace learned of Bayes's work, around 1780, Laplace acknowledged Bayes's priority and the possibility of including a factor to represent prior probabilities, but he still seldom bothered with this. In particular, he never used prior probabilities in inverse problems involving continuous variables.[7] Once he had his great theorem of 1810, he was indifferent between stating his conclusions in terms of inverse probability or stating them in Bernoullian terms; the relevant probabilities follow a normal distribution, to use Karl Pearson's name for it, and one can say this in terms of probabilities for the quantity being estimated or for the error in the estimate. There may be a small theoretical differences, but Laplace ignored them [145, 90].

Karl Pearson worked in this Laplacean tradition. In *The Grammar of Science*, the influential book on the philosophy of science he first published in 1892 [112], Pearson advised readers who wanted to learn about probability to consult Thomas Galloway's 1839 treatise [61], which taught inverse probability as developed by Laplace and Poisson, without a mention of Bayes's rule or prior probabilities. Pearson quoted Francis Edgeworth [46] 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."[8]

In the Laplacean large-sample limit, rough but solid experience might be good enough. Even if prior probabilities are unequal, why should they vary much over the small range allowed by the observations? But in Fisher's framework,

drawing with replacement from a urn containing white and black balls in unknown proportions, and Laplace used the same problem as the first example of the principle in his 1774 paper. In his painstaking study of Condorcet's unpublished work, Pierre Crépel was able to show that the manuscript was written before the spring of 1771 and perhaps even earlier ([15], pages 247–263; see also [21], pages 288–289, and [13]). It seems certain that Condorcet's example had been communicated to the younger Laplace. Prior to Crépel 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 1773, that Laplace had persuaded himself of the principle by way of a fiducial argument ([141], pages 100–101; see also [142]). It was in the theory of errors, in any case, that Laplace found the principle to have the greatest importance.

[6] As Marie-France and Bernard Bru have pointed out, this spared Laplace of any scruples about what a Bayesian would now see as using an improper uniform prior distribution [13].

[7] As Stephen Stigler has contended, hardly no one else did either until the 20th century; see [144].

[8] 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 part of Pearson's philosophy of science. He distinguished between perceptions (sense impressions) and conceptions (theories), and he saw the law of succession as fundamental to our construction of theories from repeated perceptions.

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* [80] 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 that he had learned from the *The Grammar of Science* "that scientific theories are no more than models of natural phenomena."

where the P_θ are known and theory focuses on efficiency when the number of observations falls short of the Laplacean large-sample limit, inequalities in prior probabilities do matter, and the obvious alternative to Fisher's Bernoullian methods is Bayes's rule. Little wonder that Fisher denounced this alternative so fiercely.[9]

2.2 Bayesianism

From the 1920s into the 1960s, the development of mathematical statistics was led by Fisher and then by Jerzy Neyman and E. S. Pearson, who rejected the use of Bayes's theorem in statistical analysis because of its reliance on prior probabilities. The Neyman-Pearson work was more decision-theoretic than Fisher's, and its subsequent development by Abraham Wald, Leonard J. Savage, and others, together with the development of modern game theory, brought the idea of subjective expected utility into mathematical statistics. This renewed acceptance of subjective probability led in turn to the development, beginning in the late 1950s, of a new school of thought that called itself Bayesian.[10] Influenced by decision-theoretic arguments that suggested the need for subjective probabilities, and appealing to earlier work on subjective probability by Bruno de Finetti and Frank P. Ramsey, most of the Bayesians considered it unnecessary to justify uniform probability distributions as an expression of ignorance or of rough past experience. Each person should settle on their own subjective probabilities.

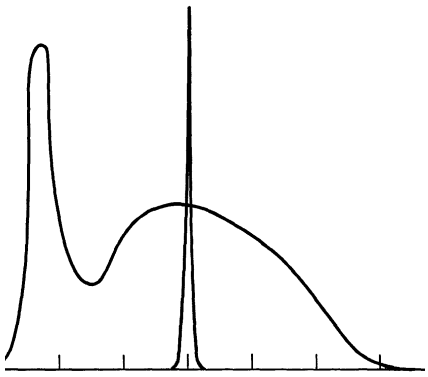
From the beginning, however, most statisticians who call themselves Bayesians, even if attracted to de Finetti's uncompromising subjective philosophy, have retained a Bernoullian interpretation of the parametric model $\{P_\theta\}_{\theta\in\Theta}$. As this implies, they want their Bayesian conclusions to have good "frequency properties".[11] When planning to announce, based on future observations, a set of values of θ that has posterior probability near one, they want to know that for each θ the probability distribution P_θ will give probability near one for observations that will result in the announced set containing θ.

[9] In his later work, Fisher convinced himself that Bayes shared his own frequency interpretation of probability and that the subjective interpretation he scorned had been introduced by Laplace [2].

[10] This story has been recounted by a number of authors; see for example [52]. The introduction of the terms *Bayesian* and *Bayesianism* is discussed in Appendix A. The use of *Bayesian* as an adjective can be found as early as 1948, but its use as a noun does not appear until the 1960s. Nor does *Bayesianism*. Before then, there were no Bayesians.

[11] Before his interaction with Savage in the 1950s, de Finetti had not worked on mathematical statistics, at least not within Fisher's parametric framework. When he first dealt with this framework, in 1953 [28], he interpreted it as representing a consensus of conditional opinion: P_θ represents the subjective probabilities everyone would have if they knew the value of θ. Such conditional opinions might be understood in a purely subjective way when θ has some reference in 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 most cases there is no such reference. In these cases, we might put a Bernoullian gloss on de Finetti's formulation: θ is the hypothesis that P_θ gives accurate frequencies or withstands gambling strategies, and you would adopt P_θ as your subjective probabilities if you knew this hypothesis is true.

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



The irregular curve is the prior density. The spiked curve is the likelihood function, which can be expected, when there many observations, to have the form of a normal density. A normal density is nearly zero at any substantial distance from its peak. So when we multiply the two curves together to obtain the posterior density, values of the product will be zero far from the peak, regardless of how large the prior density may be there.

In practice, moreover, Bayesians retained another argument with a long pedigree — the argument that the distribution of prior probabilities will not matter much so long as it is smooth.[12] In an influential article published in 1963, Ward Edwards, Harold Lindman, and Leonard J. Savage called the conclusion of this argument the *principle of stable estimation* [50]. It says that the prior probabilities required by a Bayesian analysis of a parametric model have less and less influence as the number of observations grows. The posterior probabilities for a parameter are always a compromise between the prior probabilities and the likelihood, and if the prior is smooth, the likelihood eventually dominates this compromise (see Figure 1.).

2.3 Fiducial and Dempster-Shafer arguments

Fisher always insisted that inferences about a parameter θ from observations cannot necessarily be expressed by numerical probabilities: sometimes one must stop with the likelihood function. But he was not completely insensible to the

[12] We find this argument, for example, in Arthur Lyon Bowley's widely used textbook, *The Elements of Statistics* [6], beginning with its fourth edition in 1920 (page 414). Bowley borrowed the argument from Francis Edgeworth, who made it repeatedly starting in the 1880s. See [46], [48], and the references Edgeworth gives on page 83 of the latter article.

British tradition of logical probability, which always sought a numerical probability for a given proposition on given evidence.[13] By 1930 [56], he had persuaded himself that we can sometimes obtain probabilities for θ posterior to the observations without using prior probabilities.[14] He called these probabilities *fiducial*. Suppose, to take a simple and concrete example, that $x_1, \ldots, x_{10}$ are independent and normally distributed with unknown mean μ and variance 1. Set $e = \mu - \overline{x}$, where $\overline{x}$ is the average of $x_1, \ldots, x_{10}$. Then e is normal with mean 0 and variance 1/10. Suppose now that we observe $\overline{x} = 0.23$. If we do not change our probabilities for e, then $\mu - 0.23$ is normally distributed with mean 0 and variance 1/10, and therefore μ is normally distributed with mean 0.23 and variance 1/10. In particular, the statement $-0.40 \leq \mu \leq 0.86$ has probability 95%. Later authors called e a *pivot*; its initial probability distribution does not depend on the unknown parameter μ. We may say that the argument relies on continuing to trust this probability distribution after the observations are made.

Although he offered a number of examples of his fiducial argument over the years, Fisher was never able to shape these examples into a coherent system, and by the time of his death in 1962, his fiducial project had more or less disappeared under a barrage of criticism. Eventually it became clear that fiducial probabilities did not even always have a property that Fisher himself expected of any probability. The frequency theory of probability, advanced by Richard von Mises early in the 20th century, demanded not only that the probability of an event should be its empirical frequency in repeated trials but also that there should be no way of picking out a subset of trials with a different empirical frequency. This condition follows, as von Mises explained, from the underlying meaning of probability in terms of gambling; if the condition fails one can construct a gambling strategy that will beat the probability. Apparently Fisher never mentioned von Mises in print, but in later years he adopted von Mises's condition that there be no way of picking out trials with a different frequency, calling it the absence of *recognizable subsets*. In his last book, *Statistical Methods and Scientific Inference* (1956, [58]), Fisher claimed without proof that his fiducial probabilities did not admit recognizable subsets. Shortly after his death, it was shown that this is often not true. It is not true even for Fisher's favorite example, the fiducial probabilities derived from the t-statistic for a normal mean.[15]

As originally formulated, Fisher's fiducial argument applied only to models with continuous observations, 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.[16]

[13] See Verbugt [147] for an account of the development of logical probability by De Morgan, Boole, and Jevons in 19th century Britain. For contrasting assessments of Fisher's concept of probability, see [40] and [91], pages 83–86. For a late essay by Fisher himself on the topic, see [60].

[14] For an examination of the influence of the United States statistician Mordecai Ezekiel on Fisher's 1930 paper, see [1].

[15] See [124, 151, 158] for more analysis and historical information on the fiducial story.

[16] See [59]. There are also comments in this direction in the third (posthumous) edition of *Statistical Methods and Scientific Inference* (1973).

Arthur Dempster took up the fiducial idea in a series of articles in the 1960s, giving methods for obtaining upper and lower probabilities for both continuous and discontinuous models. [30, 31, 32, 33, 35, 37, 38]. Dempster's methods constituted a generalization of the Bayesian calculus, and like the Bayesian calculus it can be used beyond the setting of parametric statistical models. I presented it in this general way in my 1976 book, *A Mathematical Theory of Evidence* [126]. In the 1980s it was widely used in artificial intelligence under the name *Dempster-Shafer theory* [157].

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.[17]

In 1982 [128], I suggested that a parametric statistical 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 that evidence, it may be possible to model it 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 [42], he has pointed that once Bayesian models and analyses are re-expressed in Dempster-Shafer terms (and thus given additional structure), it becomes clear that both the prior distribution and the likelihood function can be weakened to reflect the weakness or absence of underlying evidence.

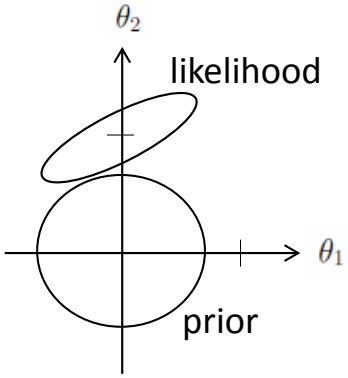
2.4 The 21st century Bayesian crisis

The Bernoullian and Bayesian theories both continue to flourish in the 21st century. Both have accommodated stochastic processes as well as random samples. But the advent of huge data sets and the concomitant complexity of models have created problems for both. By the 1980s the growth in complexity had begun to outpace Bernoullian solutions, and this, together with the development of computational methods for Bayesian analyses, allowed the Bayesian theory to mushroom in importance. But the lack of transparency in Bayesian computations with huge numbers of parameters, especially when this number far exceeds the dimensionality of the data, has now also shaken confidence in Bayesian posterior probabilities.

It has become increasingly clear that the principle of stable estimation no longer holds when the number of parameters approaches or exceeds the dimensionality of the data This is because it is impossible, before the observations are known, to provide a prior distribution that will give relatively uniform coverage,

[17] See [43] 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 [131].

Figure 2: 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.

no matter how the likelihood function comes out, over the perhaps elliptical region of values that this likelihood suggests is possible. When the dimensionality of the parameter space is orders of magnitude greater than the dimensionality of the data, as now happens frequently in fields as disparate as medicine and macroeconomics, prior probabilities can 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 [119].

Expositions of the Bayesian method for the one-dimensional case often emphasize the typical case where the both the prior and the likelihood are unimodal. In this case, we expect the posterior, since it is a compromise, to have a mode between the mode of the prior and that of the likelihood. Unfortunately, we cannot expect even this when θ is multi-dimensional. 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 $h(\theta)$, which may be of substantive interest, 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 2 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 this case of θ_2, this results in a compromise between the prior and the likelihood — a posterior mean between the prior mean 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 mean and the maximum-likelihood

estimate are both zero, but the posterior mean is negative.[18]

The example might seem contrived, but we can expect the likelihood and the prior to be tilted with respect to each other more often than not, and when this happens, and we rotate the picture so that their centers are aligned vertically, then the linear combination of θ_1 and θ_2 represented by the horizontal axis, which might be a parameter of interest, will have a posterior that is centered to one side of the vertical alignment, as in the figure. Of course, if there were so many observations that the likelihood were sufficiently peaked, then the effect would be negligible, but this is far harder to achieve in two dimensions than in one and rapidly becomes completely implausible as the number of dimensions increases. More troubling, it is increasingly difficult in high dimensions to see whether this phenomenon affects a particular feature $h(\theta)$ of interest. This problem has been studied in detail by Min-ge Xie and his colleagues; see [155] and [156], pages 27ff and the discussion with Christian Robert on pages 55, 74–75.

The failure of the principle of stable estimation has left 21st century statistical theory in a quandary. This quandary can be seen as a crisis of Bayesianism, but I will argue that it goes deeper, bringing into question not only the meaningfulness of the Bayesian prior for a model with a large number of parameters but also the meaningfulness of such models themselves.

2.5 The fiducial revival

The problems just discussed might be summarized by saying first that it is impossible, when a parameter θ has many dimensions, to provide a prior distribution that will not overwhelm 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 $\theta_1, \ldots, \theta_n$ may express very strong opinions about other features $h(\theta)$,[19] but now that we

[18] To make the example more definite, suppose 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 result could be attributed the strong prior opinion about the correlation. But there are many other ways the tilt between the likelihood and the prior could come about.

[19] 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 [138]. 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.

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, a number of statistical theorists have proposed focusing in advance on a feature $h(\theta)$ of interest and seeking posterior distributions for that feature with desired Bernoullian properties. This is hardly consistent with Bayesianism's subjectivist philosophy, and the arguments that produce the posterior distributions are often 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 approaches 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.[20] This approach was suggested by Bradley Efron in 1993 [51] and has since been developed by a number of authors, most notably Tore Schweder and Nils L. Hjort [122, 123] and Kesar Singh, Regina Liu, Min-ge Xie and their collaborators [156].

- **Generalized fiducial inference.** In this approach, developed by Jan Hannig and his collaborators [77], one identifies a *data generating equation*, $x = G(u, \theta)$, where u is an unobserved random variable with known probability distribution, x is the observation, and P_θ is the distribution for x determined by fixing θ in the equation.[21] The data generating equation is chosen to focus on the feature $h(\theta)$ of interest.[22] 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 [100], is also inspired by Dempster-Shafer theory; see [101]. Like generalized fiducial inference, it begins by adopting a data generating

[20] The name *confidence interval* was introduced by Jerzy Neyman in an effort to give a purely Bernoullian rationale and interpretation to what he thought Fisher was doing with his fiducial intervals, but the idea is much older. The idea of confidence intervals appears in Laplace's work [90], and Cournot explained the idea very clearly [16]. It also appears in the work of several early 20th century authors, from some of whom Fisher clearly drew his inspiration [1].

[21] The data generating equation is a richer structure than the parametric model it determines. In some cases, such as the venerable model error model $x = \theta + u$, this additional structure might be seen as a response to the desire for information about the evidence for the parametric model that I voiced in 1982 [127].

[22] For example, Hannig has suggested to me that in Stein's example, described in Footnote 19 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.

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 data generating 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 (non-data-dependent) prior distributions that will give them as Bayesian posteriors.[23]

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,[24] let us say that a probability becomes fiducial when we decide to trust it even though we have information not taken into account in its creation. We have decided, in other words, that the additional information is materially irrelevant. I will call this judgement of irrelevance a *fiducial judgement*.

Once we adopt this broad sense of *fiducial*, we must recognize that all probabilities are fiducial when we put them to use. We create probabilities from theory, from conjecture, or from experience of frequencies. But there is always other information.[25] To use the probabilities in a meaningful way, we must make the judgement that this other information is not materially relevant, and

[23] 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, as in Footnote 18. Some authors, such as George E. P. Box [7], have defended an iterative process of Bayesian calculation, model checking, and adjusting the prior.

[24] Initially, in his 1930 article, Fisher suggested that fiducial probabilities are probabilities of a different kind. But he 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.

[25] Permit me to deny, without repeating arguments I have made elsewhere (in [129], for example), the claim that a rational person should have already integrated all of his or her information and can find the resulting probabilities by examining his or her dispositions to act.

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.

A closer look at the historical origins of the Bernoullian and Bayesian theories will help us see their fiducial character more clearly.

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 choose a non-zero distance from p and a probability falling just short of one, then we can find a value of n such that y/n with 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. As Aleksandr Aleksandrovich Chuprov wrote to commemorate its two hundredth anniversary [110], "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 $\delta > 0$, the event

$$\left| \frac{y}{n} - p \right| \leq \epsilon \tag{1}$$

has probability at least $1 - \delta$ when n is large enough. 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 (1) 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 (1)? 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 the high probability for (1). This is a fiducial judgement. Similar judgements are required when we use the many generalizations of Bernoulli's theorem.

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 model in which the probabilities in a parametric model 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.

In Thomas Bayes's time, however, 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 $P(B|A)$,[26] should satisfy

$$P(A\&B) = P(A)P(B|A), \tag{2}$$

where $P(A)$ and $P(A\&B)$ are the initial probabilities for A and $A\&B$, respectively. The equality (2) has long been called the *rule of compound probability*. It implies, of course, that

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

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.[27] 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

$$\frac{P(A\&B)}{P(B)}. \tag{3}$$

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 (3) 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 [127]. 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

[26] I hasten to repeat that De Moivre had no such notation.

[27] 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.

explained centuries later, that this additional information does not change our attitude towards certain bets.[28]

Were we to accept Bayes's rule (3) for changing our probability for an earlier event after being told about the happening of a later event, and were we then to adopt uniform prior probabilities for the unknown prior probability p, then we could derive Bayes's formula for Bernoulli's problem of estimating p from y happenings in n trials:

$$\text{posterior probability that } a \le p \le b = \frac{\int_a^b p^y(1-p)^{n-y}dp}{\int_0^1 p^y(1-p)^{n-y}dp}. \tag{4}$$

But in an introduction to Bayes's essay, his friend Richard Price tells us that Bayes feared that his readers would not find this argument convincing and therefore gave a different argument, involving a billiard table.

The billiard 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 [23]. 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 (4) follows.

The fiducial judgement is salient here: it is 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.

Price had suggested that Bayes resorted to his second argument because readers might not agree with the assumption of a prior uniform distribution for p. The second argument, however, also uses this assumption. Perhaps it was instead the argument for his fifth proposition that Bayes found shaky. In any case, both arguments involve a fiducial judgement.

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.[29]

[28] In terms of de Finetti's picture: we do not change our disposition to make certain bets. In terms of the game-theoretic picture developed in my 2001 book with Vladimir Vovk [135]: we do not change the judgement that these bets will not allow us to multiply the capital we risk by a large factor.

[29] My translation of a passage in Section 89. See [132] for additional translations from Cournot.

The fiducial principle is another way of saying this: 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's generalization of Bayes

The arguments by Laplace, Bayes, and Crofton that we have just reviewed can all be placed within Dempster-Shafer theory and generalized in various ways. Dempster's first article on the theory included a generalization of the Bayes/Crofton argument in which we do not put prior probabilities on p and hence obtain only upper and lower posterior probabilities for it [31]. In [40], Dempster explained how the simple fiducial example I discussed on page 7 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 *Demspter'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 Bayesian arguments), 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.[30]

3.4 Imprecise and game-theoretic probability

The fiducial judgement, the judgement that we should continue to respect certain probabilities, might be applied to only some initial probabilities rather than to an entire probability distribution. We do in this in the case of Bayes's rule of conditioning and (if renormalization is required) in the more general case of Dempster's rule of combination. The new methods discussed in Section 2.5 also require this move.

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

[30]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" ([36], page 244). He now emphasizes a logical conception of probability not based on betting, in the tradition of De Morgan, Boole, and Jevons [41, 147], 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.

perhaps more plausible as a representation of actual evidence. This may take us outside Fisher's parametric framework and into the realm of imprecise and game-theoretic probability [3, 135]. For an application of the fiducial idea to the theory of imprecise probability, see [134]. For a yet more general picture in which different probability judgements are trusted to different degrees, see [70].

4 Poisson's principle

Siméon Denis Poisson (1781–1840) was Laplace's successor as the leader of French mathematics [10]. We can trace back to his work in the 1830s the principle that probabilistic prediction is possible even when probabilities vary.[31]

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 a very considerable number 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.[32]

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.[33] 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 [14]. Many other generalizations followed.

Let us first recall some of the generalizations for the simple case of coin tossing. 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

[31] And perhaps even farther back. See Chapter 9 of the second book of Laplace's *Théorie analytique* [87].

[32] [115], 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."

[33] See Stephen Stigler's summary on page 182–186 of [141]. Stigler regards I. J. Bienaymé as Poisson's most effective critic. See also [12, 78].

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 for n sufficiently large,

$$\left| \frac{\sum_{i=1}^{n} x_i}{n} - p \right| \leq \epsilon \tag{5}$$

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 for n sufficiently large,

$$\left| \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} p_i}{n} \right| \leq \epsilon. \tag{6}$$

with probability at least $1 - \delta$

Version 3 (Lévy). Suppose P is a probability distribution for $x_1, \ldots, x_n$. Then for n sufficiently large,

$$P\left(\left| \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} E(x_i | x_1, \ldots, x_{i-1})}{n} \right| \leq \epsilon \right) \geq 1 - \delta, \tag{7}$$

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 was first clearly understood by the French mathematician Paul Lévy (1886–1971) in the 1930s. 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.

All three versions 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 (5) is x's mean; p_i in (6) is x_i's mean. The conditional expected value in (7) is no longer necessarily a conditional probability.

Poisson's principle, as I have formulated it, reminds us that probability models that do not involve independent identically distributed trials can make predictions. The prediction

$$\left| \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} E(x_i | x_1, \ldots, x_{i-1})}{n} \right| \leq \epsilon \tag{8}$$

in (7) is a case in point, especially when the x_i are not binary. It does not equate a 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* [108], 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* is 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 [135], 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,

$$\overline{P}\left(\left|\frac{\sum_{i=1}^n x_i}{n} - \frac{\sum_{i=1}^n m_i}{n}\right| \le \epsilon\right) \le 1 - \delta, \tag{9}$$

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. This is hardly news. As soon as we saw the probability statement (7), we understood that that the stochastic process represented by P was predicting the event (8). But the principle needs to be stated explicitly. Cournot did so, and he was the first to state that this is the *only* way that probability relates to phenomena [136, 130].

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) [102]. It was the French mathematician Maurice Fréchet who first called this bridge *Cournot's principle* [136].

In addition to providing part of the foundation of Bernoullian statistics, Cournot's principle also helps bring Bernoullian statistics together with Bayesian statistics, 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 ([7], 1980). See also [62, 125, 133].

Most of the continental mathematicians who studied mathematical probability in the first half of the twentieth century subscribed to Cournot's principle in one way or another. Salient examples include Evgeny Slutsky, Paul Lévy, Emile Borel, Andrei Kolmogorov, Abraham Wald, and Trygve Haavelmo [102, 136, 132]. Like Chuprov, these mathematicians saw Bernoulli's theorem and its generalizations as fundamental to probability, and while this made them sympathetic with the idea that probability theory is about frequencies, they saw clearly that only one of the probabilities in Bernoulli's theorem can be equated with a frequency. The probability p in (1) is equal for practical purposes to the frequency y/n, but the probability that p is within ϵ of y/n is certainly not a frequency. Should we try to interpret it as a frequency by imagining that the whole experiment involving a long sequence of trials is itself repeated many times? This is silly at best; as Arthur Dempster pointed out in 1968, it leads straightaway into an infinite regress ([36], page 33).

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. Another classical prediction, for example, is the law of the iterated logarithm, which concerns the rate at which a frequency should converge to a probability or an average probability in repeated trials [84, 148, 135].

It is often objected against Cournot's principle 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.[34] 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 [150]. Even in theory we can only make a countable number of predictions, which could be combined into a single prediction were it computable

[34]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 [132] and Borel's discussion in his 1914 book, *Le Hasard* [5]. Neyman cited this discussion by Borel as an inspiration for the ideas in his work with E. S. Pearson on hypothesis testing [109, 89]. 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 [135] 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 [132].

[4].

Cournot's principle can be understood in terms of betting. The prediction that a particular event of small probability will not happen can be expressed by betting against that event. If the event were to happen, you would multiply the money you risked by a large factor. So the empirical meaning of a probabilistic theory comes down to the prediction that you will not multiply the money you risk by a large factor betting at the odds given by the theory. The requirement that the prediction be made in advance in implicit here, because you must make the choice in advance in order to make the bet. This betting interpretation is developed in my book with Vovk [135] and in subsequent work at `www.probabilityandfinance.com`.

The betting version of Cournot's principle extends to analyses that use Dempster-Shafer theory or imprecise probabilities, and also perhaps to analyses using other frameworks that weaken the standard probability calculus. In these frameworks the analysis of a problem may specify fewer or weaker betting offers than a full probabilistic analysis, but combinations of the offers made may still give large payoffs relative to risk for some events, and the prediction of the analysis will be that these events will not happen.

6 Summary

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, Benoullian 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.

All three principles are needed not only in Bernoullian and Bayesian analyses but also in analyses that use Dempster-Shafer belief functions or imprecise probabilities.

A Towards a history of the words *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?

A.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 ([24] 1900) and in the multiple editions of his authoritative probability textbook ([25] 1903). In the German edition of Andrei Markov's textbook, published in 1912 [99], we find both *Formel von Bayes* and *Bayesschen Formel*. In his book on the philosophy of probability ([26] 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.[35] 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 [49].[36] The French, who became remarkably disinterested in and even hostile towards Laplace's work on probability during the second half of the 19th century [13, 103], continued to use *règle de Bayes* and similar phrases. The Germans being much less interested in Laplace than in Gauss, *Bayessche* suited them well.

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 [153]. 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*

[35] 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 [54], 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".

[36] The French phrase *Méthode inverse des probabilités* appeared much earlier in unpublished teaching notes by Fourier; see [22].

assumption. The next appearance is in 1950, in prefaces R. A. Fisher wrote for two of his earlier papers [57].[37] In 1951, L. J. Savage writes of "modern, or unBayesian, statistical theory" [120].

As Stephen Fienberg has documented, the name *Bayesian* became standard in the 1950s [52]. Those who began using it then included I. J. Good, who had learned probability by reading Keynes and Ramsey in the 1930s (see the preface to [68]) and working with Turing in World War II, newly Bayesian statisticians such as Savage, and decision theorists in American business schools such as Harry V. Roberts and Robert Schlaifer. In 1958, Erich Lehmann, not a Bayesian, used the term *Bayesian derivation* in passing in an unpublished technical report [90]. By 1960, Roberts could write that "Bayesian statistics" was now a standard term ([117], page 26). By 1962, he could write about the sometimes-called "Bayesian revolution" ([118], page 202).

In the instances just cited, *Bayesian* was used as an adjective. Searches of Google Books, JSTOR, and similar databases do not turn up uses of *Bayesian* as a noun or uses of the noun *Bayesianism* before the 1960s. As I argued in Section 2.2, Bayesians and Bayesianism did not really exist until the middle of the 20th century. There was a debate starting at least in the middle of the 19th century, with Cournot, about the validity and scope of Bayes's rule, but hardly anyone contended that it provided a basis for all of mathematical statistics.

Counterparts for the newly coined English *Bayesian* and *Bayesianism* eventually came into use in other European languages. Bruno de Finetti, though Italian, was apparently the first to use the adjective *bayesien* in French, in an article published in 1955 [29]. This 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 the *Bayesianism* became *Bayesianismus*.

A.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:[38]

- 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" [47].
- Arthur Dempster advocated the usage in 1966 [31]. In 1968 [34], in a review of three volumes of collected papers by Neyman and Pearson, Dempster wrote

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

[38] *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.

> 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* [71]. 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 individual 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* [65]. In a 1956 book review, he used *orthodox* and *classical* in the same paragraph ([66], page 389). In a 1958 article, he used *classical objectivistic statistics* [67]

- Denis V. Lindley used *classical statistics* in a 1964 article [93]. In the preface to a 1965 book [94], he used *orthodox statistics*.

- John W. Pratt, in a 1965 article entitled "Bayesian interpretation of standard inference statements" [116], 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 [113] 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 [154]. Joseph L. Hodges and Erich Lehmann were still using *classical* in this way in 1961 [79].
- For many in the mid 20th century, the treatment of inverse probability by Bayes and Laplace was classical. In the chapter on confidence regions in his 1946 book [20], 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 [81, 82], Donald A. Darling in 1958 [27], and John W. Tukey in 1961 [146].
- 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 [63].

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 began in the 1950s. He did so repeatedly. 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* [121], 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 [88] and by Jaroslav Hájek in 1967 [74]. 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 [107].

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 [91], 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 [105, 106]. Nagel used *frequentist* only as a noun; the Harvard University Professor Donald Williams used it as an adjective and also used *frequentism* in 1945 [152]. 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 [149]. This framework was cumbersome compared with the axiomatics advanced by Fréchet and Kolmogorov [86], and by the end of the 1930s mathematicians had decisively rejected it as a starting point for mathematical work [136].

The term *frequentist* was first used to refer to Bernoullian statistics only in 1949, by the statistician Maurice G. Kendall [83], 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 [109], he explicitly embraced the cognomen *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 [73, 72].

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 [95]. There Lindley uses "orthodox sampling theory description", "classical sampling theory methods", "sampling theory approach", and simply "sampling theory".

In a article published in 1971 [39], 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 [98], 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" [96] and in a number of later publications (e.g., [97]). 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* [8]. In his well known 1980 discussion paper at the Royal Statistical Society ([7] 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* [19], 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 [18], 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 [64], quantum mechanics [9], and cosmology [139].

A.3 Fiducial

At the beginning of his 1873 essay on determinism [104], 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 obviously imperfect. What Fisher was taking as fixed can only be fixed as a matter of 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.

B Acknowledgements

In preparing this paper, I have benefited from conversations with Art Dempster, Andrew Gelman, Shelly Goldstein, Jan Hannig, Barry Loewer, Ryan Martin, Teddy Seidenfeld, Stephen Senn, Steve Stigler, Volodya Vovk, Min-ge Xie, Sandy Zabell, and other participants of the Fourth Bayesian, Fiducial, and Frequentist Conference, held at Harvard in May 2017 `http://statistics.fas.harvard.edu/bff4`.

References

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

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

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

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

[4] 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.

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

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

[7] 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.

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

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

[10] 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 [11].

[11] Bernard Bru. Poisson, the probability calculus, and public education. *Electronic Journal for History of Probability and Statistics*, 1(2), November 2005. Translation of [10].

[12] 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.

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

[14] 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.

[15] 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.

[16] 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 [17].

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

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

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

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

[21] 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.

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

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

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

[25] 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.

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

[27] 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.

[28] 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 du Centre national de la recherche scientifique*, pages 49–63. C.N.R.S., Paris, 1953.

[29] 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.

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

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

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

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

[34] 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.

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

[36] 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.

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

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

[39] 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.

[40] 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.

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

[42] 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.

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

[44] 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.

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

[46] Francis Y. Edgeworth. The philosophy of chance. *Mind*, 9:223–235, 1884.

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

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

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

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

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

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

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

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

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

[56] Ronald A. Fisher. Inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 26(4):528–5353, 1930.

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

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

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

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

[61] 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.

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

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

[64] 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.

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

[66] 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.

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

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

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

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

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

[72] Alan Hájek. Fifteen arguments against hypothetical frequentism. In Eagle [45], pages 410–432.

[73] Alan Hájek. "Mises redux"-redux: Fifteen arguments against finite frequentism. In Eagle [45], pages 395–409.

[74] 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.

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

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

[77] 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.

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

[79] 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.

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

[81] 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.

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

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

[84] 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.

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

[86] 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.

[87] Pierre Simon de 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*.

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

[89] 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.

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

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

[92] 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.

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

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

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

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

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

[98] 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.

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

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

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

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

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

[104] 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).

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

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

[107] 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.

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

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

[110] Kh. O. Ondar, editor. О теории вероятностей и математической статистике (переписка А. А. Маркова и А. А. Чупрова). Nauk, Moscow, 1977. See [111] for English translation.

[111] 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 [110] by Charles M. and Margaret Stein. Additional letters between Markov are provided in translation by Sheynin in [137], Chapter 8.

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

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

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

[115] 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.

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

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

[118] 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.

[119] 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*.

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

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

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

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

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

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

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

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

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

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

[130] 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.

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

[132] Glenn Shafer. Cournot in English, April 2017. GTP Working Paper 48.

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

[134] 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.

[135] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.

[136] 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*" in *Statistical Science* 21:70–98, 2006.

[137] 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.

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

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

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

[141] Stephen M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, MA, 1986.

[142] Stephen M. Stigler. Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1:359–378, 1986.

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

[144] Stephen M. Stigler. Galton visualizing Bayesian inference. *Chance*, 24(1):8–10, 2013.

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

[146] 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.

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

[148] 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.

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

[150] 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.

[151] 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.

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

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

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

[155] 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.

[156] 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.

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

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