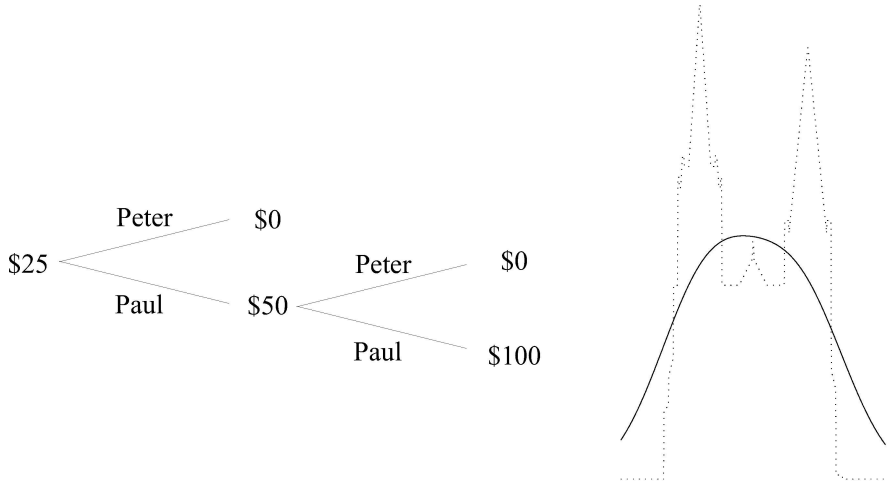
Bayesian, Fiducial, Frequentist

Glenn Shafer
Rutgers Business School
gshafer@business.rutgers.edu



The Game-Theoretic Probability and Finance Project

Working Paper #50

First posted April 30, 2017. Last revised April 30, 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 classical statisticians, Bayesians, and proponents of fiducial and Dempster-Shafer methods.

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 classical and Bayesian statisticians and to reconcile their philosophies with more venturesome ideas provided by Fisher's fiducial argument and its descendants, including the Dempster-Shafer theory of belief functions. See for example the Fourth Bayesian, Fiducial, and Frequentist Conference, held at Harvard in May 2017 `http://statistics.fas.harvard.edu/bff4`.

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*: In use, all probability is fiducial. We always have other information. To use a probability, which begins as a subjective or purely theoretical betting rate, we must make the judgement that this other information is 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.

- *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 aside 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 classical and Bayesian statisticians and other philosophers of probability talking past each other.

The fiducial principle captures a feature common to classical and Bayesian statistical practice that brings them both closer to fiducial thinking than the rhetoric of either suggests is possible. Poisson's principle and Cournot's principle can dispel the mistaken idea that classical statistical theory is based on an equation of probability with frequency. Instead, it is based on concepts of prediction and testing often shared by Bayesian practice.

Classical statistics can be traced back to Jacob Bernoulli's theorem on the estimation of a probability, published in 1713 (see Section 3.1 of this paper), but its ideas are now associated with the twentieth-century statisticians R. A. Fisher and Jerzy Neyman.[1] The name itself — *classical statistics* — peaked in use around 1960.[2] Now we more often see *frequentism* or *frequentist statistics*. In [57] and in Section 5 of this paper, I argue that *frequentism* is a misleading name for classical statistics. If we agree that *classical statistics* is no longer

[1] As Erich L. Lehmann wrote in the first sentence of his *Fisher, Neyman, and the Creation of Classical Statistics*, posthumously published in 2011 [40], "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)."

[2] According to `https://books.google.com/ngrams`.

a usable name because Bayesian statistics now appears equally classical, then I suggest the name *Cournotian statistics*. It was Antoine Augustin Cournot, in his 1843 book *Exposition de la théorie des chances et probabilités* [8], who first clearly disentangled the non-Bayesian theory that Bernoulli, De Moivre, Laplace, and others had developed from Laplace's equally innovative Bayesian theory.

2 Fiducial probability

R. A. Fisher introduced the adjective *fiducial* into statistics in 1930 [26, 67]. For decades afterwards, he used it with reference to a limited set of examples consisting of models and corresponding inferential analyses.[3] The probabilities he dubbed fiducial in these analyses aimed to answer the same questions as the posterior probabilities in Bayesian analyses of the models, but in his opinion they possessed an objectivity not achieved by the Bayesian probabilities. Fisher was never able to shape his 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.

The models in Fisher's examples all involved continuous observations, but in 1957 Fisher suggested that something similar could be done with discrete models such as the binomial, even if this did not produce precise probabilities [27, 55]. A. P. Dempster took this idea up in a series of papers in the 1960s, showing how upper and lower posterior probabilities can be obtained for both discrete and continuous parametric models [13, 14, 15, 16, 18, 19]. His 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* [51]. In the 1980s it was widely used in artificial intelligence under the name *Dempster-Shafer theory* [66]. Its current use is greatest in domains where statistical models have little traction because it is impossible, impractical, or implausible to model in advance the evidence we might obtain, and yet we want to quantify and formally combine the evidence we do obtain, including evidence that provides little or no support for either side of some of the questions being considered.[4]

Despite continued advocacy by Dempster and others, Dempster-Shafer theory has gained little traction during the past 30 years in domains where statis-

[3] Here is one of the simplest examples. Suppose $x_1, \ldots, x_{10}$ 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. Now we observe $x_1, \ldots, x_{10}$, and we see that $\overline{x} = 0.23$. If we decide not to change our probabilities about e, then we will say that $\mu - 0.23$ is normally distributed with mean 0 and variance 1/10, and therefore that μ is normally distributed with mean 0.23 and variance 1/10. In particular, we will assign 95% probability to the event $-0.40 \leq \mu \leq 0.86$. In this example, e is the *pivot*; its initial probability distribution is fully known — i.e., does not depend on the unknown parameter μ.

[4] For references, see [21] and the web site for the Belief Function and Applications Society, `http://www.bfasociety.org/`. For an accounting of my work on Dempster-Shafer belief functions in the 1970s and 1980s and its relation to my own later work, see [55].

tical models are used. But the advent of huge data sets and the concomitant complexity of models have created problems for the theories that are dominant in these domains, the classical and Bayesian theories. The Bayesian theory has mushroomed in importance, as models have outpaced classical solutions and as Bayesian computational methods have been developed. But the advent of models with huge numbers of parameters has also shaken confidence in Bayesian posterior probabilities. The comforting notion that the opinion expressed by prior probabilities will usually be swamped by data, which is plausible when there is only one parameter [23], cannot be sustained when there are many, because a prior probability distribution will necessarily encode strong opinions about some features of a multi-dimensional parameter. Min-ge Xie has pointed out to me that problems arise even for a two dimensional parameter $\theta = (\theta_1, \theta_2)$. We can often find a function $h(\theta)$, which may actually be a feature of substantive interest, such that h's posterior probabilities are not even a compromise between its prior probabilities and its likelihood[5]

In response to this conundrum, some authors have been studying fiducial or Dempster-Shafer posteriors that behave well for particular features because they are based on pivots that target those features. See the recent book by Ryan Martin and Chuanhai Liu [42] and the recent review article by Jan Hannig and his collaborators [33].

3 The fiducial principle

According to the Oxford English Dictionary, the adjective *fiducial* means "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, 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 this 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.[6] To use the probabilities in a meaningful way, we must make the judgement that this other information is not materially relevant, and this makes the probabilities fiducial. This is just as true for classical and Bayesian probabilities as it is for the fiducial probabilities that Fisher invented.

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

[5] See [64] and [65], page 27ff and the discussion with Christian Robert on pages 55, 74–75.

[6] Permit me to deny, without repeating arguments I have made elsewhere (in [53], for example), the claim that a rational person should have already integrated all of his or her evidence and can find the resulting probabilities by examining his or her dispositions to act.

3.1 Bernoullian estimation

If an event with probability p happens y times in n independent trials, and n is sufficiently large, then y/n will be arbitrarily close to p with probability arbitrarily close to one. This is Jacob Bernoulli's theorem, first published in 1713. It is justly celebrated. As Alexander Alexandrovich Chuprov wrote to commemorate its two hundredth anniversary, "everywhere the logic of inference rests in the final account on the theorem of Jacob Bernoulli."

Here is a slightly more formal statement of the 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 the theorem in any particular case is legitimatized by the judgement that the additional information (y and the very fact that we have learned y) 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 from a parametric statistical 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.

When Pierre Simon de Laplace and Thomas Bayes first introduced what we now call the Bayesian method, however, there was no such thing as conditional probability — no such concept, no formal definition, and certainly no notation for it. Thus a "Bayes's theorem" was not possible. There could only be a "Bayes's formula" or a "Bayes's rule". The ways in which Laplace and Bayes introduced this rule betray their fiducial thinking.

3.2.1 Laplace's inverse probability

In the 1774 paper in which he introduced Bayes's rule, Laplace gave no argument for the rule.[7] He simply stated it as a principle:

> If an event can be produced by a number n of different causes, then the probabilities of the existence of these causes taken from the event are to each other as the probabilities of the event taken from these causes, and the probability of the existence of each of them is equal to the probability of the event taken from that cause, divided by the sum of all the probabilities of the event taken from each of these causes.[8]

Forty years later, in his philosophical essay on probabilities, Laplace again stated this principle without proof, adding only an explanation of how to handle unequal prior probabilities:

> Each cause to which an event can be attributed is indicated with more likelihood to the extent that it is more probable that the event would have happened had that cause existed. So the probability of the existence of one of the causes is a fraction in which the numerator is the probability of the event resulting from that cause and the denominator is the sum of the corresponding probabilities for all the causes. If these various causes, considered *a priori*, are unequally likely, then the probability of the event resulting from each cause should be replaced by its product with the probability of the cause itself.[9]

How did Laplace arrive at his principle of inverse probability? After studying a paper Laplace drafted earlier but never published, Stigler concluded that Laplace had made a fiducial mistake while thinking about random errors in astronomical observation ([61], pages 100-101):

---

[7] By all accounts, Laplace was not aware of Bayes's 1763 essay at this time. It came to the attention of Laplace's colleague Condorcet and presumably also to Laplace's attention a couple of years later [61, 12].

[8] Laplace 1774 [38], page 623, my translation. French original: "Si un événement peut être produit par un nombre n de causes differentes, les probabilités de l'existence de ces causes prises de l'événement sont entre elles comme les probabilités de l'événement prises de ces causes, et la probabilité de l'existence de chacune d'elles est égale à la probabilité de l'événement prise de cette cause, divisée par la somme de toutes les probabilités de l'événement prises de chacune de ces causes." Stigler ([61], page 102) translates "prises de", which literally means "taken from" as "given"; this helps us relate the passage to our contemporary concept of conditional probability but might also mislead us into thinking that Laplace had this concept.

[9] Laplace [39], page xiv, my translation. French original: "Chacune des causes auxquelles un événement observé peut être attribué est indiquée avec d'autant plus de vraisemblance qu'il est plus probable que, cette cause étant supposée exister, l'événement aura lieu; la probabilité de l'existence d'une quelconque de ces causes est donc une fraction dont le numérateur est la probabilité de l'événement résultante de cette cause; et dont le dénominateur est la somme des probabilités semblables relatives à toutes les causes: si ces diverses causes, considérées *à priori*, sont inégalement probables; il faut, au lieu de la probabilité de l'événement résultante de chaque cause, employer le produit de cette probabilité par celle de la cause elle-même."

> Once a random distribution of errors was conceived ... the inversion could follow almost inadvertently. If e represents the error, O the observation, and P the point observed, then $O = P + e$ implies equally well that $P = O - e$. If e is taken as randomly and symmetrically distributed, then supposing P fixed gives a distribution for O; and conversely, taking O as given leads to a distribution for P. By treating only the *difference* $e = O - P$ as random, a symmetrical situation is created where inversion becomes most natural. R. A. Fisher was to call this a fiducial argument .... Statisticians now know that this argument is not so simple as a naive view might have it, that deep and subtle philosophical and mathematical points must be dealt with if this most natural approach is to lead to a coherent theory of inference. But in the mid-eighteenth century the argument was a conceptual liberation. Whether expressed in mathematical symbols or in words, it led naturally to a view in which one distribution — the distribution of errors — provided the random element for both "forward" (in time) and "inverse" probability statements. And it suggested an idea of inverse inference, reasoning *probabilistically* from effect (O) to cause (P), that was to bear fruit in other applications as well.

3.2.2 Bayes's fifth proposition

In the 18th century, when Bayes was studying probability, the general notion of conditional probability had not been invented. 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 $\mathbf{P}(B|A)$,[10] should satisfy

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

where $\mathbf{P}(A)$ and $\mathbf{P}(A\&B)$ 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

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

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.[11] To turn $\mathbf{P}(A)\mathbf{P}(B|A)$ into one

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

[11] 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.

monetary unit if $A\&B$ happens, you first bet it all on A; this gives you $\mathbf{P}(B|A)$ 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 $\mathbf{P}(A)$ to

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

This is the fifth proposition in Bayes's essay, but the proof was hardly a proof. Bayes 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 $\mathbf{P}(A)$ to the ratio (3) is then convincing. But this does not establish that the change makes sense in other cases, where we do have other information, and where we may not have known in advance what we would be told when, so that the very fact that are told about B without being told about A is itself information [52]. To use Bayes's fifth proposition, we must make the fiducial judgement that this additional information is irrelevant. We must decide, as Bruno de Finetti explained centuries later, that this additional information does not change our willingness to make certain bets.

3.2.3 Bayes's second argument

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:

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

We could even call this a theorem. 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 [11]. The question "will not be altered" Crofton opined, if we suppose that whether the event happens on 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.

It is notable that there was still no Bayes's theorem in 1885, because there was still no general notion of conditional probability. When Bayes's formula became Bayes's theorem is a delicate question, but a reasonable answer is 1901, when the German mathematician Felix Hausdorff proposed taking (3) as a general definition of the probability of one event given another [34].

3.2.4 Cournot's assessment of Bayes

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.[12]

The fiducial principle is another way of saying this: we should use the betting rate only if we make the judgement that other information (other than B's happening and the information that went into fixing $\mathbf{P}(B)$ and $\mathbf{P}(A\&B)$) is irrelevant.

3.3 Dempster's generalization of Bayes

I will not undertake an exposition of the Dempster-Shafer theory here, but it is worth noting that 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 [13]. In [20], Dempster explained how the simple fiducial argument that Stigler attributes to

[12] My translation of a passage in Section 89. See [56] for additional translations from Cournot.

Laplace 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 in our conversations in the 1970s, we "continue to believe". As I now prefer to say, we continue to be willing to make certain bets.[13] 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.

3.4 Imprecise and game-theoretic probability

The fiducial judgement, the judgement that we should continue to believe certain probabilities or continue to offer certain bets, 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. *Confidence distributions* for individual parameters also require this move [65].

If we anticipate that we might retain only certain probabilities, it is reasonable to ask whether some of those that we will not retain can be identified at the outset and removed from the initial model, thus making this model simpler and perhaps more plausible. This can bring us to the logic of imprecise and game-theoretic probability [2, 59]. For an application of the fiducial idea to the theory of imprecise probability, see [58]. For a more general picture in which different probability judgements are trusted to different degrees, see [29].

4 Poisson's principle

Siméon Denis Poisson (1781–1840) was Laplace's successor as the leader of French mathematics [5]. We can trace back to his work in the 1830s the principle that probabilistic prediction is possible even when probabilities vary.

4.1 The law of large numbers

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

[13] Dempster has seldom discussed the connection between probability and betting, but he once observed that "the connection is so close that it is almost of the nature of a tautology to speak of one or of the other" ([17], page 244).

> 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.[14]

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 perhaps Bernoulli's theorem, applied to the mean probability, is theory enough.) 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 [7]. Many other generalizations followed.

To fix ideas, let us 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 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,

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

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,

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

with probability at least $1 - \delta$

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

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

[14] [49], page 478, my translation. The original French: "Les choses de toute nature sont soumises 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."

where $\mathbf{E}(x_i|x_1,\ldots,x_{i-1})$, the expected value under $\mathbf{P}$ 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 case, 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. All the three versions generalize to the case where the random variables $x_1,\ldots,x_n$ are not necessarily binary but satisfy certain regularity conditions. We replace p with x's mean μ in (5) and p_i with x_i's mean μ_i in (6). No change is required in the expression (7), but the conditional expected value is no longer necessarily a conditional probability.

Version 3 was first clearly understood by the Russian mathematician Sergei Bernstein (1880–1968) in the 1920s and 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 late 1930s, as they more fully absorbed the Russian and French thinking as a result of the influx of continental mathematicians fleeing Hitler.

The law of large numbers is further generalized game-theoretically in [59], 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. In this case, assuming for example that the x_i and m_i are all bounded in absolute value by the same constant C, we find that for n sufficiently large

$$\overline{\mathbf{P}}\left(\left|\frac{\sum_{i=1}^n x_i}{n} - \frac{\sum_{i=1}^n m_i}{n}\right| \le \epsilon\right) \le 1-\delta, \tag{8}$$

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

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

$$\left|\frac{\sum_{i=1}^n x_i}{n} - \frac{\sum_{i=1}^n \mathbf{E}(x_i|x_1,\ldots,x_{i-1})}{n}\right| \le \epsilon \tag{9}$$

in (7) is a case in point.

4.2 The ubiquity of stochastic processes

In one sense, 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. In 1960, in the *Journal of the American Statistical Association* [47], 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.

Even earlier, Trygve Haavelmo had convinced econometricians that probability theory is not limited to the picture of sampling from a population that had been popularized by Karl Pearson and R. A. Fisher. In his 1944 article, "The probability approach to econometrics" ([30], pages 477–478), Haavelmo wrote:

> The reluctance among economists to accept probability models as a basis for economic research has, it seems, been founded upon a very narrow concept of probability and random variables. Probability schemes, it is held, apply only to such phenomena as lottery drawings, or, at best, to those series of observations where each observation may be considered as an independent drawing from one and the same 'population'. From this point of view it has been argued, e.g., that most economic time series do not conform well to any probability model, 'because the successive observations are not independent'. But it is *not* necessary that the observations should be independent and that they should all follow the same one-dimensional probability law. It is sufficient to assume that the *whole set* of, say n, observations may be considered as *one* observation of n variables (or a 'sample point') following an n-dimensional *joint* probability law, the 'existence' of which may be purely hypothetical. Then, one can test hypotheses regarding this joint probability law, and draw inferences as to its possible form, by means of *one* sample point (in n dimensions). Modern statistical theory has made progress in solving such problems of statistical inference.

Haavlemo's argument is seen by historians as a decisive step towards modern econometrics [24, 44, 50, 35, 1].

5 Cournot's principle

To put Poisson's principle to work, we must of course 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 $\mathbf{P}$ was predicting the event (9). But the principle needs to be stated explicitly. Cournot was the first to do so.

If we agree with Chuprov that classical 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). It was the French mathematician Maurice Fréchet who first called this bridge *Cournot's principle* [60].

In addition to providing part of the foundation of classical statistics, Cournot's principle also helps bring classical statistics together with Bayesian statistics, because thoughtful Bayesian statisticians also believe in model checking. In the end, a Bayesian model is of little use in practice unless its probabilistic 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 ([4], 1980) and the recent article by Andrew Gelman and Cosma Rohilla Shalizi [28]. See also [57].

5.1 More on the history of Cournot's principle

Here is how Cournot stated the principle in his 1843 book:

> *The physically impossible event is therefore the one that has infinitely small probability*, and only this remark gives substance— objective and phenomenal value—to the theory of mathematical probability.[15]

For additional translations from Cournot, including passages from his discussion of significance testing in general and multiple testing in particular, see [56].

In his 1944 article, Haavelmo stated Cournot's principle in a more modern way ([30], page 478):

> The class of scientific statements that can be expressed in probability terms is enormous. In fact, this class contains all the 'laws' that have, so far, been formulated. For such 'laws' say no more and no less than this: The probability is almost 1 that a certain event will occur.

The continental mathematicians who studied mathematical probability in the first half of the twentieth century almost all explicitly subscribed to Cournot's principle in one way or another. Salient examples include Evgeny Slutsky and Andrei Kolmogorov in Russia and Paul Lévy and Emile Borel in France [43, 60, 54]. Kolmogorov included a statement of Cournot's principle in the celebrated monograph in which he gave the modern axioms for probability [37]. Like Chuprov, these mathematicians saw Bernoulli's theorem and its generalizations as fundamental to probability, and while this made them sympathetic in a certain sense to a "frequency interpretation" of probability, 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

[15] [8], 1843, Section 43. The orginal French: "*L'événement physiquement impossible est donc celui dont la probabilité mathématique est infiniment petite*; et cette seule remarque donne une consistance, une valeur objective et phénoménale à la théorie de la probabilité mathématique."

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 Dempster pointed out in 1968, it leads straightaway into an infinite regress ([17], 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 [62, 59].

It is sometimes objected against Cournot's principle that an event with small probability always happens: a lottery always has a winning ticket. This overlooks the role of the statistician or scientist, who chooses the prediction in advance. Injecting an observer into the picture might seem to threaten the objectivity of the probability model, but in practice only a limited number of predictions are important [63]. Even in theory we can only make a countable number of predictions, which could be combined into a single prediction were it computable [3].

5.2 The name *frequentism* is misleading.

Once we make Cournot's principle explicit, we are in a position to appreciate the unsuitability of the names *frequentism* and *frequentist*. They suggest a naive equation of probability with frequency that hardly does justice to classical statistics. By embracing these words, classical statisticians have driven many philosophers to conclude that the Bayesian paradigm has no coherent competition [31, 22].

How did *frequentist* become attached to classical statistics? By all accounts, the name was coined by the American philosopher Ernest Nagel (1901–1985) in 1936 [45, 46]. The first statistician to use it was Maurice G. Kendall (1907–1983), in 1949 [36], 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 [48], he explicitly embraced the cognomen *frequentist*.

R. A. Fisher's continuing influence has also played a role. Anders Hald has written that Fisher "was a genius who almost single-handedly created the foundation for modem statistical science" ([32], page 738), and Fisher put the notion of a random sample from a population at the foundation of this foundation. To see this, it suffices to recall a few sentences from Fisher's immensely influence 1922 article, "On the mathematical foundations of theoretical statistics" [25]:

> In order to arrive at a distinct formulation of statistical problems, it is necessary to define the task to which the statistician sets himself: briefly, and in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which

> usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information in the original data.
>
> This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all qualities under discussion. Any information given by the sample, which is of use in estimating the values of these parameters, is relevant information. ...

In spite of the importance of stochastic processes, Fisher's random sample often still holds a ghostly grip on the imagination of mathematical statistics. It dominates statistical teaching at the elementary and not-so-elementary level. We might even count it as a marker distinguishing the culture of mathematical statistics from the culture of mathematical probability.

Some statisticians who see themselves as heirs of the non-fiducial aspects of Fisher's thought subscribe to what they call the *repeated sampling principle*. This principle has been stated by David R. Cox and David V. Hinkley in these words ([10], page 45):

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

Is this principle necessary or even appealing when we are dealing with complex stochastic processes for which repetition is impossible? Was Haavelmo mistaken to think that we can test a model and make inferences and predictions from the data we actually have? Was Neyman mistaken to think that we can rely on a stochastic process after confirming that it explains observed frequencies and averages? What, aside from Fisher's ghost, forces us to imagine nearly unimaginable hypothetical repetitions? In my view, Cournot's principle frees classical statistics from any need for the repeated sampling principle. Classical statistics should refuse to be called *sampling-theory statistics*.[16]

5.3 Let's rename classical statistics after Cournot.

In the introduction, I suggested that we call classical statistics *Cournotian*.

The name *Bayesian statistics* has come to be accepted not because Bayes contributed more than Laplace to the development of Bayesian statistics, but because Laplace's probability was not exclusively Bayesian. We no longer use the nineteenth-century term *inverse probability* because *inverse* is often beside the point.

[16]The Bayesian statistician Dennis Lindley was fond of calling it by this name; see for example [41].

Cournot stands in a similar relation to classical statistics. He did not make stellar mathematical contributions as Karl Pearson, R. A. Fisher, and Jerzy Neyman did, but by this very token his name is not associated with any of the views that separate one of these giants from the others. He stands nearer beginning, and he can be counted as a founder inasmuch as he first identified and defended the classical paradigm that Laplace had inherited and developed, removing from it the Bayesian elements that Laplace had mixed in so freely. He had a clear view of significance testing, and he may have been the first to think carefully about multiple testing [56, 57]. Classical statistics is Cournotian.

References

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

[1] John Aldrich. The econometricians' statisticians, 1895–1945. *History of Political Economy*, 42:111–154, 2010.

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

[3] 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, 2009.

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

[5] 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 [6].

[6] Bernard Bru. Poisson, the probability calculus, and public education. *Electronic Journal for History of Probability and Statistics*, 1(2), November 2005. Translation of [5]. `http://www.jehps.net/`.

[7] 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. `https://eudml.org/doc/183251`.

[8] 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 [9].

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

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

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

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

[13] Arthur P. Dempster. New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, 37:355–374, 1966.

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

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

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

[17] 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.

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

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

[20] 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.

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

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

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

[24] Roy J. Epstein. *A History of Econometrics*. North-Holland, Amsterdam, 1987.

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

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

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

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

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

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

[31] Alan Hàjek. Fifteen arguments against hypothetical frequentism. In *Philosophy of Probability: Contemporary Readings*, pages 410–432. Routledge, 2011.

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

[33] 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.

[34] Felix Hausdorff. Beiträge zur Wahrscheinlichkeitsrechnung. *Sitzungsberichte der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse*, 53:152–178, 1901.

[35] James J. Heckman. Haavelmo and the birth of modern economics: A review of *The History of Econometric Ideas* by Mary Morgan. *Journal of Economic Literature*, 30:876–886, 1992.

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

[37] 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.

[38] Pierre Simon de Laplace. Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Académie royale des sciences de Paris*, 6:621–656, 1774. Reprinted in Volume 8 of Laplace's *Oeuvres complètes*, pages 27–65.

[39] Pierre Simon de Laplace. *Essai philosophique sur les probabilités*. Courcier, Paris, first edition, 1814. The fifth and definitive edition was published in 1825. A modern edition, edited by Bernard Bru, was published by Christian Bourgois, Paris, in 1986. An English translation of the fifth edition by Andrew I. Dale was published by Springer in 1995.

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

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

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

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

[44] Mary Morgan. *The History of Econometric Ideas*. Cambridge University Press, Cambridge, 1990.

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

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

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

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

[49] 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. `http://gallica.bnf.fr/ark:/12148/bpt6k29606/f473.image.langEN`.

[50] Duo Qin. *The Formation of Econometrics: A Historical Perspective*. Oxford, New York, 1993.

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

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

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

[54] 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.

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

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

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

[58] 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.

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

[60] 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.

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

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

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