Statistical Inference from a Dempster-Shafer Perspective: What has Changed over 50 Years?

A. P. Dempster

Department of Statistics, Harvard University

Workshop on the Theory of Belief Functions

ENSIETA, Brest, France, April 1-2, 2010

Abstract: In the first part of my talk, I review background leading up to the early DS inference methods of the 1960s, stressing their origins in R. A, Fisher's "fiducial argument". In the second part, I survey present attitudes, describing DS outputs as triples (p, q, r) of personal probabilities, where p is probability for the truth of an assertion, q is probability against, and r is a new probability of "don't know". I describe DS analysis in terms of independent probability components defined on margins of a state space model, to which computational operations of projection and combination are applied. I illustrate with a DS treatment of standard nonparametric inference, and an extension that "weakens" to gain protection against overly optimistic betting rules. Finally, I suggest that DS is an appropriate framework for the analysis of complex systems.

Circa 1950, I was an undergraduate studying mainly mathematics, and some physics, in a strong program at the University of Toronto. Circa 1955, I was a PhD student in the Princeton Mathematics Department, having decided to go and study mathematical statistics with Sam Wilks and John Tukey. It was a good program, blending British applied mathematical emphasis with the more abstract Berkeley-based mathematical emphasis on what Jerzy Neyman called *inductive behavior*, all together with American pragmatism concerning applications.

Circa 1960, I found myself in the young Harvard Statistics Department that had been founded by Fred Mosteller. Through teaching mainly PhD students, I started to really learn and understand the field, with specific interests mainly in multivariate sampling theory and theories of statistical inference. I started out knowing probability from the side of mathematical probability, and statistics mainly from the standpoint of the Berkeley-Columbia-Stanford frequentist school. I was exposed to many applications especially in a wide range of social and biomedical sciences, and was familiar with the nascent Bayesian group at the Harvard Business School.

I then started reading and reflecting on the long history of probabilistic inference, from its 17th Century beginnings, through Bernoulli and Laplace, and most especially, up to and including R. A. Fisher. Fisher was near the end of a remarkable career that kick-started 20th Century statistics into becoming the profession and academic discipline it is today, first by deriving many basic sampling distributions, and then by introducing and naming basic inferential concepts, including notions and theory related to parameter estimation and to significance testing. Many of his new technical concepts, among them sufficiency, likelihood, and efficiency, remain basic. His concept of randomized experimentation remains a "gold standard" in many areas of empirical research. All in all, it was an exciting time to be entering the field.

The active debates in the 1950s lay in the friction between the Neymanian “frequentists”, and the reawakened “Bayesians” represented by L. J. “Jimmie” Savage (USA) and Dennis Lindley (UK), and also by Howard Raiffa and colleagues at the Harvard Business School. Fisher was out of the loop, generally viewed as part of history, regarded as a genius, but often wrong. In the profession, surprisingly little has changed. Fisher died while visiting Adelaide in 1962, before I ever encountered him personally, but I was fortunate to attend an obscure lecture by him in a crowded hall at Columbia University, while he was passing through on his way to Australia.

I found Fisher intriguing, because he rejected both of the popular movements, and because I resonated to his reasoning. He believed, as I do, that the primary role of mathematics is to support science, in his case by providing tools for valid reasoning about specific uncertainties due to sampling error and to observational noise. Thus although he was the source of many applications of sampling distributions, he rejected the Jerzy Neyman's emphasis on long run properties of procedures, seeing these to be removed from immediate practical needs of scientific inference, and overly mathematical. By 1960 he had few friends in the mathematical statistics community.

Through much of his career Fisher criticized the use of "inverse probability", as the Bayesian scheme was called a century ago. Since he gave the impression that his science was meant to be objective, entirely divorced from subjective elements, and because he worked almost entirely with sampling distributions, he has been regarded by frequentist statisticians as a founding father of their theory, despite his strong objections to their central idea. Toward the end of his life, he seems to have mellowed somewhat toward Bayesian ideas, but not to my knowledge toward Neyman and what he regarded as the frequentist corruption of statistical science.

I recount all this because in 1930 Fisher introduced, through the example of how to assess the correlation coefficient of a Gaussian bivariate population, what he called “fiducial” inference, which is a direct antecedent of DS inference. Despite severe criticism, he continued to provide new illustrations, and he held fast for the remainder of his life to claims that fiducial probabilities are valid, specifically, that fiducial posterior probabilities can, with caveats and restrictions, be obtained without the use of Bayesian prior distributions. This is also the claim of DS models and analyses, which, although not in full agreement with Fisher’s ideas, have much in common with fiducial.

Logical difficulties and controversies over fiducial inference could be topics for several lectures. Bayesians complained about violations of what they called “coherence”. Frequentists maintained that Fisher had made a simple mistake and refused to own up to it. Fiducial examples were condemned as fundamentally nonunique. An opposite sin of omission, recognized by Fisher himself , was that the fiducial paradigm could not handle discrete observables, even when sample data are simple dichotomies, as had been analyzed already in the 18th Century by Bernoulli and Bayes, using their innovative treatments based on sampling distributions, and conditional posterior distributions, respectively.

Circa 1960, I started to ponder how fiducial reasoning might be extended to the classical example of sampling a dichotomous population. Results of this thinking appeared in three basic papers in 1966, 1967, and 1968, the first two in the *Annals of Mathematical Statistics*, and the third a discussion paper in the *Journal of the Royal Statistical Society Series B*. The first of these laid out the earliest instance of what I now call a DS model and analysis, the goal being the same as that addressed by Bernoulli and Bayes, and with a proposed extension from sampling a dichotomy to sampling a polytomy with any number of categories. The second paper laid out the abstract framework of DS inference as I still understand it, and as it is explained in more detail in Glenn Shafer's pathbreaking 1976 monograph.

The third paper explained how Bayesian models and analyses are special type of DS models and analyses, requiring the strong assumption of a joint prior pdf of all the variables in the model. When the simple DS model of my 1966 paper is applied without a Bayesian prior distribution, the need for upper and lower posterior probability inferences pops up. These were not available to Fisher, whence his inability to include discrete data in his fiducial canon. In effect, weakened Bayesian assumptions led to weakened inferences, now recognizable as all within the DS framework.

I wrote several other related papers in the 60s, and supervised several Harvard PhD theses by students, all of whom went on to successful careers that made no formal use of DS ideas. The 1960s were inauspicious for DS, in part because the DS examples developed in those years referred to situations already served reasonably well by older methodologies, and in part because computations required by proposed DS methods were much more intensive than those required by Bayesian methods, which themselves surpassed the capabilities of 1960s computers. Both of these factors have now greatly changed. Models currently under development do address issues not amenable to Bayesian treatment, and the computing scene is vastly different.

In the remainder of this talk, I make no attempt to review developments from 1970 to 2000. Instead, I jump ahead to describe a current view DS inference, and assess prospects for DS becoming well established on the main stage of contemporary statistical sciences. These prospects depend on more than laying out the formal mathematics of the DS system. First, the mathematics needs to be accompanied by a sense of how DS can usefully add capabilities to everyday uncertainty assessment by working professionals and decision-makers. There also need to be convincing illustrations of these capabilities, both simple and complex.

Fundamental to establishing DS methodology as a player in real world uncertainty assessment is the notion of personal probability. There is a hurdle here for my statistical colleagues, because teaching in the statistical professions has mainly assumed that probability models are primarily idealized mathematical representations of randomness. Does probability connote personal assessment of uncertainty, or does it merely convey a property of objective randomness? I maintain that from its earliest origins much of the intuition and motivation behind the mathematics of probability includes the former. Opposition to the personalist viewpoint has been strong, however, especially from mathematical theorists.

Bruno de Finetti famously remarked that probabilities do not exist. Perhaps he might equally have remarked that randomness does not exist. Randomness is largely an illusion covering the inability of human analysts to understand and represent highly complex systems that underlie even the simplest of phenomena. Personal probabilities demonstrably do exist, however, because DS and Bayesian analyses produce them following well-defined rules of procedure, and real world analysts use them to guide choices. Surely a well-rounded view of science requires accepting inputs both from objective real world sources together with informed subjective judgment. Exclusion of one or the other is an untenable position.

There is a serious distinction behind the decision by Jimmie Savage circa 1960 to stop using the term subjective probability in favor of personal probability. Use of the latter is not just a matter of substituting less controversial language for the same troublesome concept. A personalist, as I understand the term, is quite comfortable with obtaining personal probabilities from objective real world sources, such as a large sample relative frequency. There is, moreover, more than just an observed frequency involved, beginning with the necessity of a personal judgment of equal weights among the units being counted, and often continuing with the application of personally selected statistical analyses. The situation is thus asymmetric between the two attitudes. The personalist approach is inclusive, while objectivists try to exclude personalism.

With these issues out of the way, at least for now, we can proceed to defining and illustrating the DS system. There are two distinct parts to a DS model, that I will refer to as the state space model (SSM) and DS probability models (DSPMs). It is fundamental that these constructions rest on logically distinct efforts. Each draws on a different kind of understanding of the situation under analysis. Both are nontrivial in actual practice. I am moved to issue these warnings, because a paper in mathematical statistics typically begins by stating something like " $X_1, X_2, \ldots, X_n$ are independently and identically distributed random variables" with such and such a parametric, or semi-parametric, or nonparametric stochastic model. The presentation is so stylized as to trivialize the science involved in getting to that point, let alone understanding how the mathematics relates to any actual uncertainties.

Recently I have been suggesting that for DS analyses in practice, terms like random, or stochastic, or aleatory, should be replaced by “pepr”, standing for “personal probability”, so that for example “random variable” becomes “pepr variable”. For many decades, the *lingua franca* of mathematical discussions of probability has used terms like “random variable” and “stochastic process”. Such language encourages users of the theory to think in terms of a world “governed” or “generated” by a type of phenomenon that is random. No doubt this is illuminating. It is also a mental construct. It is not inconsistent with the broader mental construct of personal probability. Hence my proposed use of “pepr” to cover the broader interpretation.

I am getting ahead of myself, however, because pepr is a concept entering only at the stage of model construction that brings in DSPMs. Variables like X_1, X_2, … , X_n arise first in specifying the SSM. For example, I might identify X_1, X_2, … , X_n as the logarithms of 2009 gross household incomes in USD of a defined sample of American families, where the sample is a subset of a defined population whose effectively infinite size can be represented by a continuous cdf F(x). Notice that no mention of probability, pepr or otherwise, is as yet present. Having the SSM, however, you can ask the basic question confronting the statistician, "What can be inferred about the unknown F(x) given known values of X_1, X_2, … , X_n?"

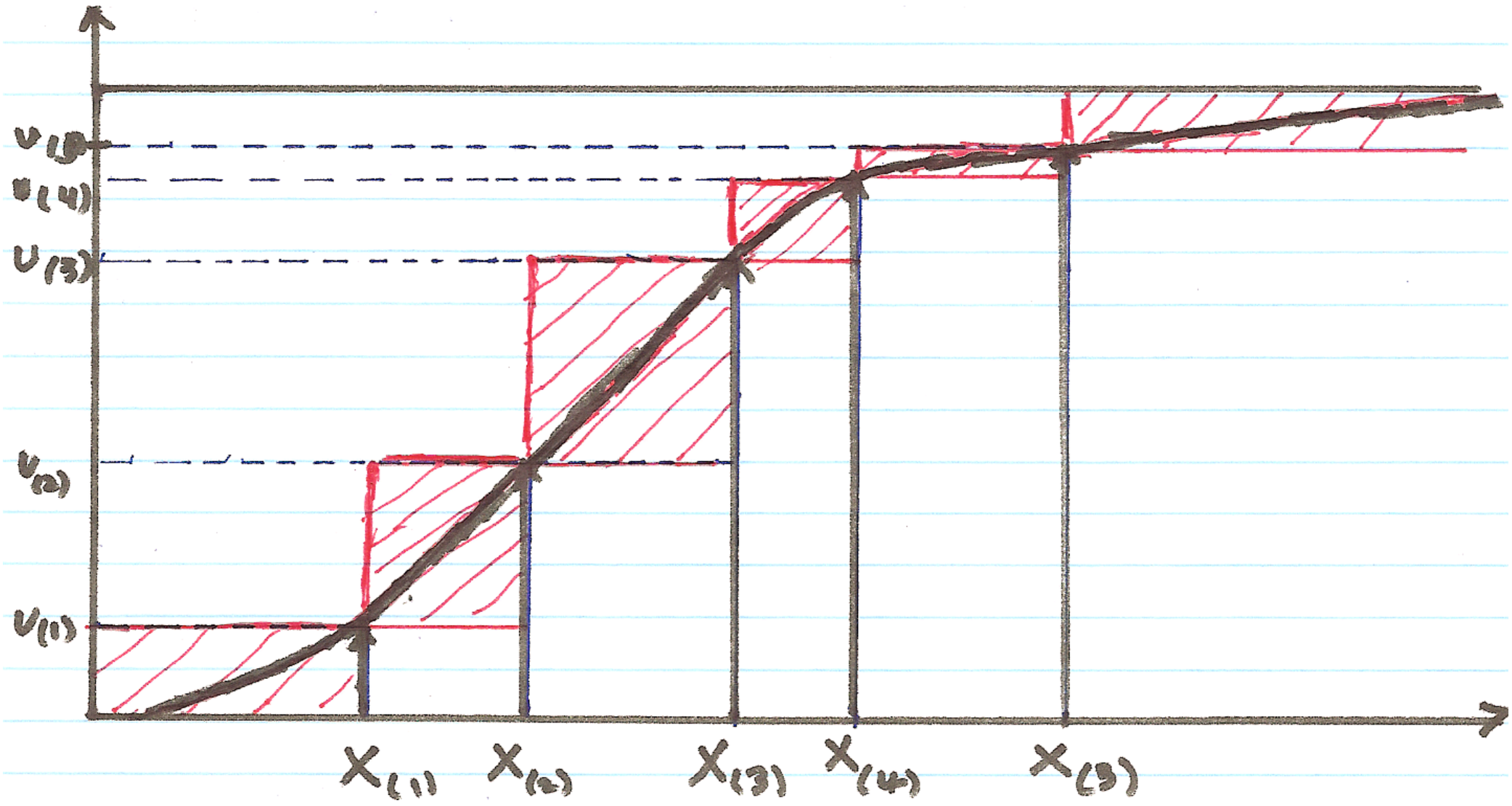
After an SSM has been defined, both mathematically as a system of variables, and in context as representing possible states of a defined real world system, then, and only then, may the analyst move on to specifying uncertain knowledge about the actual state. This is accomplished by defining a collection of independent DSPMs over the SSM. A DSPM is specified mathematically using a pepr distribution called a mass distribution over a partition of the SSM. Such a mass distribution is understood to represent evidence, generally uncertain or incomplete evidence, about where in the SSM of possible states the true state lies. Independence among members of a collection DSPMs is understood to mean noninterference among the evidential sources, in the sense that accepting the assessment from one DSPM does not alter the assessment from the others.

The meaning of DSPMs is illustrated through building on the example where the SSM consists of a sample $X_1, X_2, \ldots, X_n$ together with a continuous F(x). Since functions of the variables defining the SSM are also part of the SSM, due to having possible states and an unknown true state, we can define further variables $U_1, U_2, \ldots, U_n$ where $U_i = F(X_i)$ for $i = 1, 2, \ldots, n$. The basic building block of the DS version of nonparametric statistics is defined by creating a vector of independent and identically distributed uniform pepr variables $u_1, u_2, \ldots, u_n$, and defining a DSPM by assigning the mass distribution of $u_1, u_2, \ldots, u_n$ to the state space variables $U_1, U_2, \ldots, U_n$. This assumption was in effect first set out by Fisher in a 1939 obituary for W. S. Gosset (“Student”) , and eventually became the basis of a large literature on nonparametric statistics.

Along with the basic DSPM of uniform pepr variables, n further DSPMs are determined by observations x_1, x_2, … , x_n on the SSM variables X_1, X_2, … , X_n. These n DSPMs are of the special kind that assigns pepr one to a single numerical value. DS analysis now proceeds by projecting these n+1 DSPMs on margins up to logically equivalent DSPMs on the full SSM. Assuming independence they are then combined into a single DSPM by the algorithm that Shafer called Dempster's rule, that is actually just the mathematical definition of DS independence. The combined DSPM is the basis for making (p, q, r) evaluations of statements about the true state of the SSM.

The foregoing sketch of DS analysis is too brief to be assimilated *de novo*. Very few of my statistical colleagues have tried, let alone succeeded, mainly due I think to a mindset that excludes pepr from the realm of technical concepts. In fact, of course, the DS scheme is incredibly simple from a technical standpoint, which I believe is an essential part of its appeal. I will not try to fill in details of the mathematics, but instead proceed to a sketch of nonparametric inference, which Fisher introduced as an example of fiducial reasoning, tossing it off as an incidental extension of a claim that Gosset adopted fiducial thinking, both in the context of Student's t, and in relation to what later writers would call nonparametric tolerance regions.

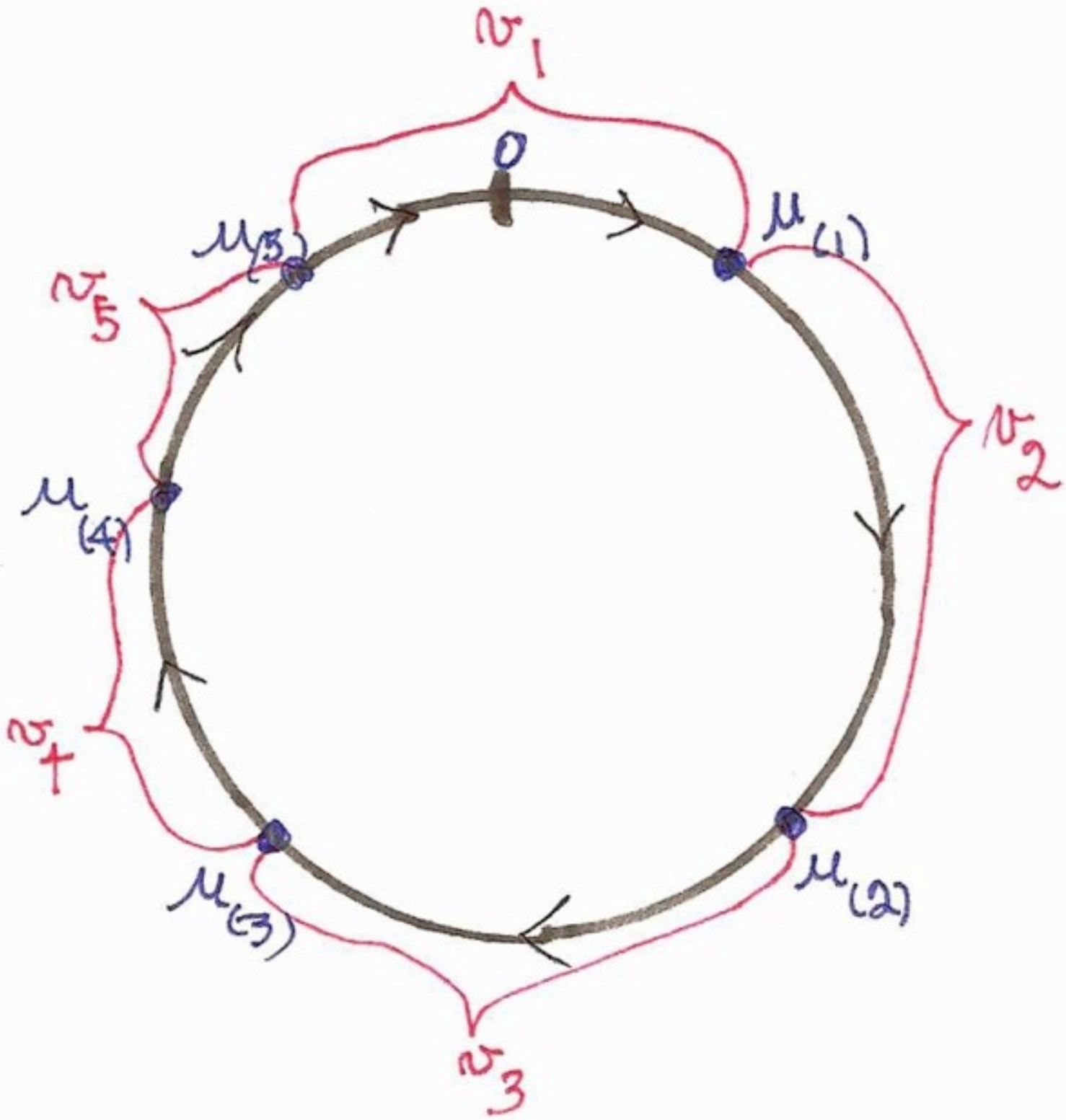
By working from the picture in Figure 1 on the next slide, I argue that Fisher was actually close to the (p, q, r) formulation of nonparametric statistical inference, but did not make the last step to the third rail of "don't know", that is required to include a much extended catalog of illustrative inferences. Then I continue by discussing a weakened version of the standard nonparametric model that illustrates a principle that I call "protective weakening", where protection refers to allowing counterparties to choose between the two sides of a bet, or more broadly among many different bets on offer. Finally I will indicate why I believe DS offers possibilities for improved analyses of complex systems.


v(5)
v(4)
v(3)
v(2)
v(1)
x(1)
x(2)
x(3)
x(4)
x(5)

The message that Figure 1 is intended to illustrate is that the values of the unknown cdf F(x) at the 5 observed values of X are DS posterior pepr distributed like 5 ordered draws from a uniform distribution on (0, 1). This inference permits DS pepr inferences of the form (p, q, 0) about many interesting unknowns, such as changes in F(x) between any pair of observed X's, or how a future sample of X's is pepr distributed among the observed X's. It also permits DS inferences of the form (p, q, r) allowing $r > 0$ for unrestricted aspects of F(x), such as F(b) – F(a) for any a and b, or for any statements about future sample X's, such as how many will be found between a and b.

It is not difficult for mathematically trained academic statisticians to write down many analytic formulas that capture expressions for (p, q, r) inferences in the example illustrated by Figure 1. Alternatively, it may be easier to think of generating desired inferences from using Monte Carlo output from a computer. It is straightforward to replicate a large number of pictures like Figure 1, using repeated pseudorandom draws of 5 uniform pepr variables. Then by recalling that F(x) is tied down only to remain in the shaded boxes, it is easy to accumulate simulated frequencies that approximate desired (p, q, r) inferences. Passing directly from pictures to algorithms bypasses the need to understand any analytic formulas.

Personal probability is often used for evaluating proposed bets. If you have computed the (p, q, r) that F(b) – F(a) exceeds .95 for some a and b, you have selected, you are permitted to offer 100p euros to win 100 euros if F(b) – F(a) turns out to exceed .95, at least if you accept that the opponent starts with the same evidence that you possess. You are advised not to allow your betting opponent to choose the side of the bet, and even more strongly advised not to allow your counterparty to select a and b as well as the side of the bet. Figure 2 illustrates a way to protect against one way that a counterparty could select against you if you do allow both choice of a and b and choice of the side of the bet.


v_1
0
$u_{(1)}$
$u_{(5)}$
v_5
$u_{(4)}$
v_2
v_4
$u_{(3)}$
$u_{(2)}$
v_3

The mathematical point is that when the 5 uniform draws illustrated in Figure 1 are regarded as plotted around a circle of circumference 1, then the pepr distribution of the arc lengths v_1, v_2, v_3, v_4, v_5 is symmetric under all rotations of the circle. The basic inference in the nonparametric scheme proposed by Fisher assumes that the arcs are placed as shown. A more protected DSPM is obtained by weakening the assumption by leaving the arc lengths as determined by the ordered uniforms as in Figure 1, but asserting that you "don't know" the angle of rotation. Your counterparty does not thereby win in the long run by choosing among patterns of X's, for example, by judging that a cluster of close X's is probably accidental. Further protection can be attained by permuting as well as rotating.

Finally, I comment that the sequential structure of the DS model construction, first a SSM, then DSPMs, permits analysis of complex real world systems that recognize a distinction that is often blurred. It is relatively easy to characterize an SSM of effectively unlimited size by specifying billions of variables that proceed through time. It is unthinkable to provide a meaningful pepr distribution over such a huge SSM, as in fact the specialization to Bayes must either attempt, or back down to assuming that most variables can be regarded as DS independent of variables in a feasible Bayesian analysis.

By contrast, if a huge panoply of variables are deemed essential to an adequate portrayal of processes underlying a complex system, then within the DS framework it becomes possible to retain all such variables in an SSM, while limiting assigned DSPMs to evidence judged credible. The result will include explicit measures of "don't know", first to be admitted, and then faced, sometimes by further scientific efforts. At present I am not aware of real world examples that illustrate this principle. It should not be difficult to develop modest illustrations taking at least baby steps beyond the limitations of the Bayesian paradigm.