False confidence, non-additive beliefs, and valid statistical inference

Ryan Martin[*]

February 3, 2019

Abstract

Statistics has made tremendous advances since the times of Fisher, Neyman, Jeffreys, and others, but the fundamental questions about probability and inference that puzzled our founding fathers still exist and might even be more relevant today. To overcome these challenges, I propose to look beyond the two dominating schools of thought and ask what do scientists need out of statistics, do the existing frameworks meet these needs, and, if not, how to fill the void? To the first question, I contend that scientists seek to convert their data, posited statistical model, etc., into calibrated degrees of belief about quantities of interest. To the second question, I argue that any framework that returns additive beliefs, i.e., probabilities, necessarily suffers from *false confidence*—certain false hypotheses tend to be assigned high probability—and, therefore, risks making systematically misleading conclusions. This reveals the fundamental importance of *non-additive beliefs* in the context of statistical inference. But non-additivity alone is not enough so, to the third question, I offer a sufficient condition, called *validity*, for avoiding false confidence, and present a framework, based on random sets and belief functions, that provably meets this condition.

Keywords and phrases: Bayes; fiducial; foundations of statistics; inferential model; p-value; plausibility function; random set.

1 Introduction

Statistics is concerned with the collection and analysis of data for the purpose of testing existing theories and creating new ones, i.e., learning about the world through observation. This is central to the scientific method and, therefore, a solid foundation of statistics is fundamental to the advancement of science. These foundations have been severely shaken in recent years, however, starting perhaps with Ioannidis's provocative claim that *most published research findings are false* (Ioannidis 2005). Science is a "culture of doubt" (Feynman 1956), so Ioannidis's skepticism is healthy. Unfortunately, the support for his claim has continued to mount—see, e.g., Nuzzo (2014), the reports in *Nature*[1] on "Challenges in Irreproducible Research," etc.—and the gravity of the problem is now becoming clear: large-scale replication projects have found that just $37/97 = 39\%$ of findings in psychology (Open Science Collaboration 2015), $11/18 = 61\%$ in experimental

---

[*]Department of Statistics, North Carolina State University. Email: `rgmarti3@ncsu.edu`
[1]`http://www.nature.com/nature/focus/reproducibility/index.html`

economics (Camerer et al. 2016), and $13/21 = 62\%$ in social science (Camerer et al. 2018) could be replicated. Rightfully so, this plight is now being called a *replication crisis*.

There are many factors that contribute to the replication crisis, some are statistics-related and others are not.[2] While statisticians can't fully resolve the replication crisis, it's important that we acknowledge how we've contributed to the problem and do our part to help. There are many ways to do so, but since the crisis is an attack on our foundations, I believe that these require our direct attention.

Sadly, it's necessary to emphasize that the foundations of statistics are practical, so let me kick off my foundational discussion in the same way as Barnard (1985):

> *I shall be concerned with the foundations of the subject. But in case it should be thought that this means I am not here strongly concerned with practical applications, let me say that confusion about the foundations of the subject is responsible, in my opinion, for much of the misuse of statistics that one meets in fields of application...*

The need to establish the practical relevance of foundations stems from the tendency for discussions of the latter to turn abstract and philosophical, blurring their connection to everyday statistical analysis. Both abstraction and philosophy are necessary but, in order to be effective, they must be applied at the right level and in the right direction. The two dominant schools of thought—*frequentist* and *Bayesian*—have co-existed for a century despite many philosophical arguments for one approach over the other. The reason these arguments have failed to firmly establish the foundations of statistics is that they generally don't align with the well-established intuition, experience, and methodology developed by data analysts. An important example is *Birnbaum's theorem*[3] (e.g., Berger and Wolpert 1984; Birnbaum 1962), which implies that the use of, say, the z- or t-test about an unknown normal mean is, in a certain sense, illogical or lacks foundational support, despite its (non-controversial) theoretical properties and many years of successful applications. Needless to say, practitioners doubt the relevance of such a result, so its impact has been rather limited.[4]

At the end of the day, the frequentist and Bayesian philosophies can co-exist because we lack the necessary understanding to say what's "right." The question isn't *frequentist or Bayesian?* but, rather, *how best to meet the needs of science?* Obviously, philosophers of science would have something to say about this, and Mayo (2018) gives a delightful, modern, and comprehensive account of how she sees us getting "beyond the statistics wars." My specific path here is different from Professor Mayo's although we have similar

---

[2]Sociological factors, such as the "publish or perish" culture of academia, are major contributors to the replication crisis; see Crane and Martin (2018a,b,d) and Section 9 below.

[3]There are two generally accepted notions in statistics: the *sufficiency principle*, which says that inferences should depend only on the value of the minimal sufficient statistic, and the *conditionality principle*, which says, roughly, inference should be conditioned on the observed values of ancillary statistics. A more controversial notion is the *likelihood principle*, which says that inferences should only depend on the shape of the likelihood function. Birnbaum's theorem states that the sufficiency and conditionality principles imply the likelihood principle. This is problematic because really only a Bayesian approach satisfies the likelihood principle. Therefore, if I use, say, a t-test, which violates the likelihood principle, then Birnbaum's theorem says I'm making an illogical step somewhere because I must have violated the well-accepted sufficiency or conditionality principles.

[4]In addition to the doubt about the relevance of Birnbaum's theorem, there has been doubt about its factuality. Indeed, Evans (2013) and Mayo (2014) have argued that the claim is wrong or at least that its implications have been exaggerated.

objectives: we both aim to satisfy that astute outsider in Fraser (2011b) who asks "Why don't [statisticians] put their house in order?"

Statisticians know what an effective statistical analysis looks like, so, whatever would be the foundations of the subject, it should be virtually uncontroversial. Towards this, I like what Reid and Cox (2015) have to say:

> ...for any analysis that hopes to shed light on the structure of the problem, modeling and calibrated inferences ...seem essential.

To set the scene, my focus throughout this paper will be on the inference aspect. This is not because modeling is easy or of lesser importance, but because the quantity about which inferences are sought is often (though not always) defined only relative to the posited model. Since the inference problem needn't exist without a model, I will assume that a model is already in place, and focus exclusively on the "calibrated inferences seem essential" part. It's still uncontroversial, but to make this suggestion operational, a precise understanding of what each word means is necessary.

- *Inference.* While not strictly necessary, there is a strong desire to quantify uncertainties via probabilities or, more generally, via degrees of belief, and to make inferences based on these. Moreover, though it'll take us a while to get there, I'll argue in Section 8.4 that my choice to formulate inferences based on degrees of belief is, in a certain sense, without loss of generality.

- *Calibrated.* In metrology, calibration refers to the comparison of measurements to a specified scale. In the present context, it's the data analyst's degrees of belief that require a scale for interpretation, e.g., what belief values are sufficiently "large" or "small" that justified conclusions can be drawn. Another side of calibration, which appears in a follow-up quote from Reid and Cox (see Section 4 below), is that the scale on which beliefs are interpreted should prevent "systematically misleading conclusions," i.e., provide control on the frequency of errors.

- *Essential.* This word in the trio I take very seriously, for reasons that I will explain. Academic statisticians are encouraged (and incentivized) to focus on "methodology," the development of methods and software to be used (hopefully widely) by practitioners and scientists. Therefore, the system is set up so that the experts, the ones who develop methods, are involved in data analyses only at the inference step, and even that involvement is indirect through the practitioner's use of their method/software. If this is the extent of statisticians' involvement in most scientific applications, should we be satisfied with methods that provide "approximate calibration" for "most hypotheses"? I say no, so here I'll take "essential" literally, that calibrated beliefs are absolutely necessary.

Despite what this introduction might suggest, this paper is not really about foundations. In particular, I make no attempt to write out any formal, rigorous, foundational system. The above discussion was simply to establish that the foundations of statistical inference are intended to serve practitioners so, whatever these foundations might look like, they should align with the goals of practitioners. That is, if "calibrated inferences seem essential" for an effective statistical analysis, as I believe they are, then the foundations ought to lead practitioners in that direction. An alternative, more realistic goal is that if a particular framework or method does not necessarily lead practitioners in the desired direction, then the foundations should at least weed this one out. The main foundational contribution of this paper is in this latter vein.

After some setup and notation, I introduce main object of study in Section 2, namely, an *inferential model*, a map from data, statistical model, etc., to a function that returns the data analyst's degrees of belief in hypotheses of interest. This formalizes the basic idea that statistics is about the conversion of data and other inputs into beliefs about the world. These inferential models can take many forms, but the most common are those that return *additive* degrees of belief, i.e., probabilities. In Section 3, I argue that additive inferential models are always at risk for being miscalibrated, and that this can and does lead to "systematically misleading conclusions" (Reid and Cox 2015). Therefore, whatever the foundations of statistical inference might be, if they are to align with the "calibrated inferences seem essential" standard, then they ought to disavow any approach that makes inferences based on probabilities. This suggests that there is something fundamental about non-additive beliefs in this context. The presence of non-additivity can be felt in the logic of classical statistical inference—e.g., rejecting the null hypothesis does not mean accepting the alternative—but, to my knowledge, the extent of non-additivity's importance has yet to be fully elucidated.

This observation about additivity is just the beginning of the story, since non-additivity alone is not enough to ward off false confidence. Additional constraints on the inferential model output are needed, beyond non-additivity, and Section 4 presents what I call the *validity condition*, a mathematical formulation of the requirements contained in the phrase "calibrated inferences seem essential." After some important remarks in Section 5, and a review of non-additive beliefs in Section 6, I describe the framework from Martin and Liu (2016) that is able to meet the validity condition. What distinguishes this approach from others is the incorporation of a suitable, user-specified random set, which results in a provably valid inferential model that returns a genuinely non-additive belief function. General details about the construction along with a number of examples are presented in Section 7. After another collection of remarks in Section 8, e.g., comparing this valid inferential model to Fisher's fiducial argument, I conclude in Section 9 with some perspectives, open problems, etc.

2 Inferential models

2.1 Setup, notation, and objectives

To start, I assume that there is a statistical model $\mathsf{P}_{Y|\theta}$, indexed by a parameter $\theta \in \Theta$, for the observable data $Y \in \mathbb{Y}$. Here, as is common in discussions of inference, I will assume that both the data and model are given, but they both deserve a few comments; see also Crane and Martin (2018c).

- Intuitively, data is what the statistician gets to see; in fact, the data is the only thing in a data analysis that's really known. Data must certainly be relevant to the scientific problem, and the more informative the better. Technical details about the collection of (relevant and informative) data are presented in, e.g., Hedayat and Sinha (1991) and Hinkelmann and Kempthorne (2008). Here I use the adjective "observable" to distinguish the data Y from other random variables that are "unobservable;" see Section 7.

- Lehmann and Casella (1998, Ch. 1) define a model as a family of probability measures on $\mathbb{Y}$, indexed by a parameter $\theta \in \Theta$:

$$\mathscr{P} = \{\mathsf{P}_{Y|\theta} : \theta \in \Theta\}. \tag{1}$$

(Here $\mathsf{P}_{Y|\theta}$ is a probability distribution for Y, a random variable, random vector, or whatever, on $\mathbb{Y}$, depending on θ; it need not be interpreted as a conditional distribution of Y, given θ, derived from a joint distribution for the pair.) Guidelines on the development of sound models are given, e.g., in Cox and Hinkley (1974), Box (1980), and McCullagh and Nelder (1983); McCullagh (2002) gives a more technical discussion. Here I will assume that the statistical model (1) is derived from some knowledge about the data-generating process and/or from an exploratory data analysis. While it's not a focus of the present paper, this procedure of developing sound models is far from routine, especially in modern, complex problems such as network analysis (e.g., Crane 2018d).

Some might say that the model is incomplete in the sense that it's missing an uncertainty assessment about the parameter θ in the form of a prior; see, for example, de Finetti (1972), Savage (1972), Gelman et al. (2004), and Kadane (2011). If a real subjective prior is available, then the statistical inference problem is "simply an application of probability" (Kadane 2011, p. xxv), in particular, Bayes's theorem (e.g., Efron 2013a) yields the conditional distribution of θ, given $Y = y$, which can be used for inference. If genuine prior information is available, then surely one should use it. However, a substantial part of the literature adopts a view that no real prior information is available; even in the Bayesian literature, the trend is to work with "default" or "non-informative" priors that let the data speak for itself. One reason for this interest in prior-free scenarios is, as Brad Efron said during a presentation at the 2016 Joint Statistical Meetings in Chicago, that "scientists like to work on new problems." In any case, my choice to omit a prior from the model is to explore what can be done in cases where a fully satisfactory prior distribution is not available. Incorporating prior information into the framework I'm describing here is possible and I'll have more to say on this in Section 9.

Given a sound statistical model and relevant data, the goal is to make certain statements about the unknown θ. Classically, these "statements" come in the form of point or interval estimates, hypothesis tests, etc.; prediction could also be included in this list of tasks (e.g., Martin and Lingham 2016), but my focus here is on inference about fixed-but-unknown quantities. A selling point of Bayesian and other distribution-based approaches to statistical is that it is relatively straightforward to read off, for example, point and interval estimates, from the posterior distribution. So, in the same spirit, I will take this "posterior distribution" or, more generally, the data analyst's degrees of belief based on data and posited statistical model as the primitive, with estimates and tests being derivatives. The next subsection makes this more precise.

2.2 Definition

In the Bayesian literature (e.g., Kadane 2011; Lindley 2014), those elements discussed in the previous section are converted to a probability distribution on Θ that encodes the data analyst's degrees of belief, from which inferences will be drawn. In other words, the Bayesian procedure, or *model*, for inference is one in which the various inputs—relevant data, sound statistical model, etc—are mapped to probabilities concerning the unknown quantities. This idea can be generalized, as in the following definition, allowing for degrees of belief output that need not be a probability distribution.

Definition 1. Fix the sample space $\mathbb{Y}$, parameter space Θ, and let $\mathscr{P}$ be a statistical model as in (1) that connects the two. Then an *inferential model* is a map from $(\mathscr{P}, y, \ldots)$

to a function $b_y : 2^\Theta \to [0, 1]$ where, for each hypothesis $A \subseteq \Theta$ concerning the unknown parameter, $b_y(A)$ represents the data analyst's degree of belief in the truthfulness of the assertion "$\theta \in A$."

This definition formalizes the idea that inference is based on a process of converting observations into degrees of belief about θ. That is, an inferential model is simply a rule or procedure that describes how the observed data, posited statistical model, and perhaps other ingredients (e.g., a prior distribution) are processed to quantify uncertainty for the purpose of making inference.

This inferential model notion captures the familiar Bayesian approach, where the output b_y is simply the posterior distribution for θ, given data y, under the assumed model and a user-specified prior distribution. It also captures fiducial inference (Barnard 1995; Fisher 1973; Zabell 1992), structural inference (Fraser 1968), generalized inference (Chiang 2001; Weerahandi 1993), generalized fiducial inference (Hannig 2009, 2013; Hannig et al. 2016, 2006; Hannig and Lee 2009; Lai et al. 2015), confidence distributions (Schweder and Hjort 2002, 2016; Xie and Singh 2013), and Bayesian inference with data-dependent priors (Fraser et al. 2010; Martin and Walker 2014, 2017). The key feature these all share is that their inferential output, b_y, is a probability distribution. But Definition 1 does not require b_y to be a probability measure, so it also covers situations where b_y is something more general, such as a capacity, belief function, etc. In these latter approaches, b_y is not additive, so it is possible that *both* $b_y(A)$ and $b_y(A^c)$ are small if data y is not especially informative about A. It turns out that additive versus non-additive has important consequences; see Section 3.

Whatever the specific mathematical form of b_y, it can be used in a natural and straightforward way. That is, large values of $b_y(A)$ and $b_y(A^c)$ suggest that data strongly supports the truthfulness and falsity, respectively, of the claim "$\theta \in A$."[5] See Dempster (2014) for a discussion of this general approach to inference through degrees of belief. Moreover, the data analyst can, if desired, produce decision procedures based on the inferential output, e.g., reject a null hypothesis "$\theta \in A$" if and only if $b_y(A^c)$ is sufficiently large.

Lastly, it should be clear that inferential models are not unique—my degrees of belief might be different from yours. That inference requires individual judgment is not a shortcoming, it is what makes statistics an interesting subject and what makes statisticians' expertise valuable. Depending on the application, however, I may want my degrees of belief to be meaningful to others, not just to myself, which puts certain constraints on my inferential model. As Evans (2015, p. xvi) writes, "subjectivity can never be avoided but its effects can be... controlled." This control comes from external or, in some sense, "non-subjective" considerations, and the remainder of the paper focuses on these.

3   Additive inferential models

3.1   Satellite conjunction analysis

Satellites orbiting Earth impact our everyday lives. For example, those of us with a poor sense of direction rely on GPS navigation applications that use positioning information from satellites to successfully reach our destinations. But it is not enough just to get these satellites into orbit, they must be constantly monitored to prevent collision with

---

[5]Intermediate values of $b_y(A)$, such as 0.4, are more difficult to interpret, but the same is true for probabilities; see Section 5.3.

one another or with other kinds of debris. The challenge is that collision events cannot be determined with certainty, so conjunction analysts must make important decisions, e.g., whether evasive maneuvers are necessary, in the face of uncertainty. This practically-relevant application will serve as motivation for the forthcoming general discussion of additive versus non-additive beliefs.

It is standard for conjunction analysts to use probabilities for uncertainty quantification, and their approach is as follows. Consider a pair of orbiting satellites whose positions and velocities are elements in three-space. The standard conjunction analysis starts with an approximation to reduce it to a two-dimensional problem. That is, a plane of closest approach is determined and then it is assumed that the satellites are each approaching this plane from an orthogonal direction, so only the projections onto this plane are relevant. Let $\theta = (\theta_1, \theta_2)^\top$ represent the difference between the true but unknown satellite positions on this plane, so that small $\|\theta\|$ indicates the satellites are close, hence a potential collision. Let $Y = (Y_1, Y_2)^\top$ denote the estimate of this difference in satellite positions based on noisy data. A simple statistical model to describe this scenario is

$$Y \sim \mathsf{N}_2(\theta, \Sigma).$$

The normality assumption is often justified by invoking the central limit theorem, since the estimate Y is the result of aggregating many position and velocity measurements. Here, as is common in the conjunction analysis literature, I will assume that the covariance matrix Σ can be accurately estimated so it will be taken as fixed.

The next step in the analysis, as reviewed in Balch et al. (2017), is to first construct a corresponding posterior distribution for θ. Since θ is a location parameter, one might consider an inferential model that describes degrees of belief based on

$$(\theta \mid y) \sim \Pi_y := \mathsf{N}_2(y, \Sigma).$$

This is a flat-prior Bayes posterior and a fiducial/confidence distribution. Using information on the satellites' sizes, a collision radius, t, can be derived, so that a collision event is defined as "$\|\theta\| \leq t$." The analysis proceeds by calculating the probability of collision (or non-collision) based on the posterior distribution above. While experienced statisticians may have reasons for concern (see below), this approach is not obviously unreasonable.

However, an interesting and paradoxical phenomenon has been observed in the conjunction analysis literature, called *probability dilution*. The basic idea is that lower data quality (i.e., more noise in the position and velocity measurements) implies lower estimated collision probability. This is paradoxical because lower quality data clearly means more uncertainty but the probability calculation gives a different conclusion, namely, that there is more certainty in the satellite's safety. Of course, if conjunction analysts take a low collision probability at face value, then they are likely to conclude that the satellite is safe when, in fact, it may not be. To see this clearly, consider the simple scenario where $\Sigma = \sigma^2 I_2$. Then the non-collision probability, according to the aforementioned posterior distribution, is a simple non-central chi-square calculation:

$$p_y(\sigma) := \Pi_y(\|\theta\| > t) = 1 - \mathtt{pchisq}(t^2/\sigma^2, \mathtt{df} = 2, \mathtt{ncp} = \|y\|^2/\sigma^2).$$

Figure 1 plots this non-collision probability as a function of σ. When σ is small, there is no problem; but when σ is large, the non-collision probabilities are large, *no matter what the data y says*. That is, even if the data suggests a collision, like in the $\|y\|^2 = 0.5$ case,
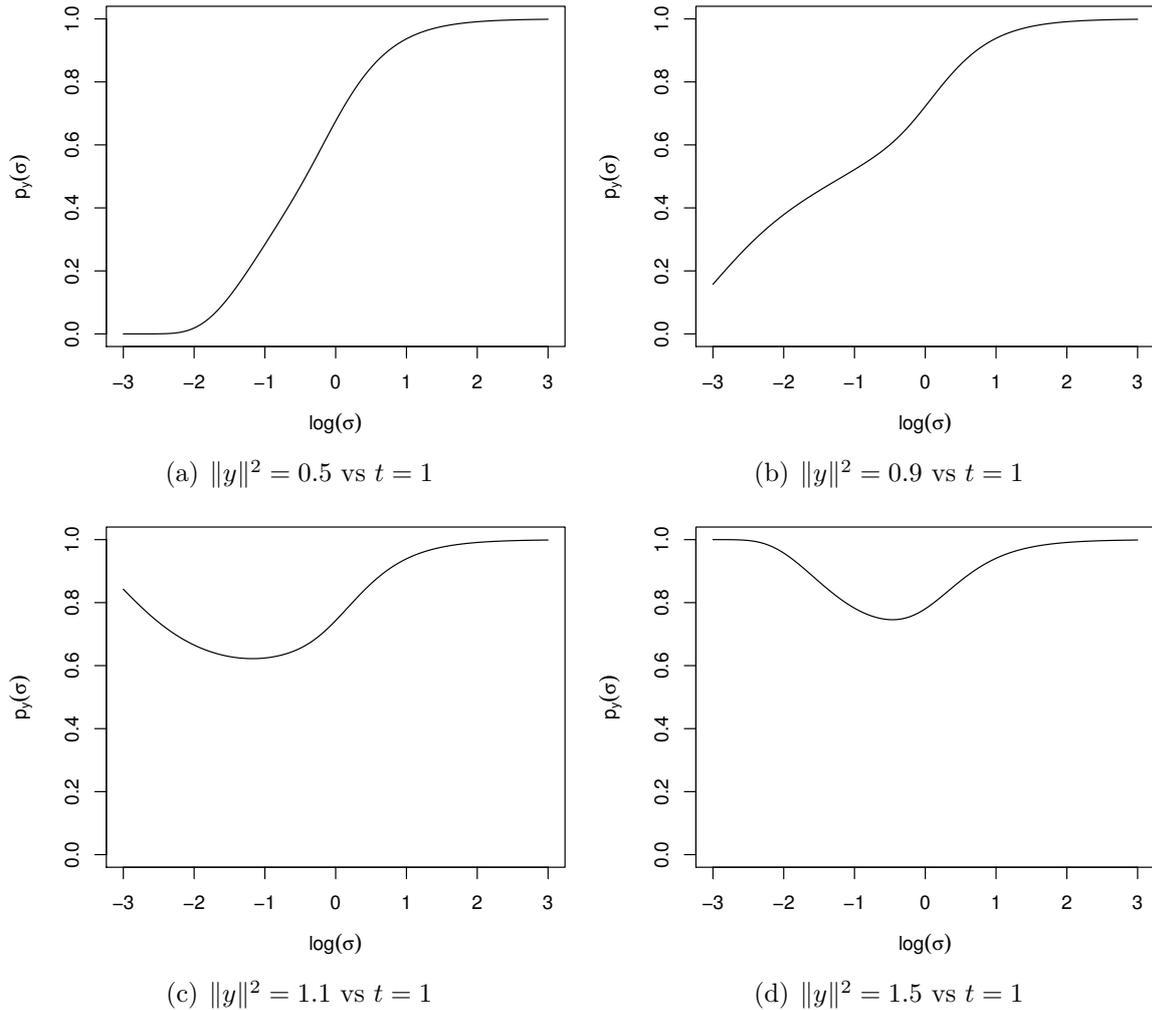
(a) $\|y\|^2 = 0.5$ vs $t = 1$

(b) $\|y\|^2 = 0.9$ vs $t = 1$

(c) $\|y\|^2 = 1.1$ vs $t = 1$

(d) $\|y\|^2 = 1.5$ vs $t = 1$

Figure 1: Plots of the non-collision probability, $p_y(\sigma)$, as a function of σ for several configurations of y. Here the collision threshold is $t = 1$.

with enough noise, the posterior probabilities indicate that the satellite is safe. This turns out to be a special case of the *false confidence* phenomenon discussed in Sections 3.2–3.3.

Statisticians will recognize this problem—inference on the length of a multivariate normal mean vector—from Stein (1959), so the paradoxical conclusions might be expected. But this normal means problem is not the only one where issues like this arise, so a better understanding of the causes and effects is valuable. Besides, the insights gleaned from Stein's investigation of this deceptively simple problem, in particular, the notions of *shrinkage* and *borrowing of strength*, are now the foundations of modern high-dimensional data analysis, so it is worth digging into these peculiar examples further.

3.2   Normal coefficient of variation

For data $Y = (Y_1, \ldots, Y_n)$, consider a statistical model that posits the components of Y to be independent and identically distributed (iid), with $Y_i \sim \mathsf{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$. Here $\theta = (\mu, \sigma^2)$ is the unknown parameter, but primary interest is in $\phi = \sigma/\mu$, a quantity called the *coefficient of variation*. One possible approach is to construct a Bayesian

inferential model where, given a prior distribution for θ, the corresponding posterior for θ is obtained via Bayes's formula, and then a marginal posterior for ϕ is derived from that of θ using the ordinary probability calculus.

Since interest is in ϕ, I will adopt the reference prior for θ, now expressed as $\theta = (\phi, \sigma)$, presented in Berger et al. (1999, Example 7), designed specifically for cases where inference on ϕ is the goal. Using the standard default conditional prior $\pi(\sigma \mid \phi) \propto \sigma^{-1}$, for σ, given ϕ, and the reference prior for ψ,

$$\pi(\phi) \propto \{|\phi|(\phi^2 + \tfrac{1}{2})^{1/2}\}^{-1},$$

given in their Equation (38), one obtains the proper marginal posterior for ϕ:

$$\pi_y(\phi) \propto \pi(\phi) e^{-\frac{n}{2\phi^2}(1-\frac{\bar{y}^2}{D^2})} \int_0^\infty z^{n-1} e^{-\frac{nD^2}{2}(z-\frac{\bar{y}}{D^2\phi})^2} \, dz, \tag{2}$$

where $\bar{y} = n^{-1}\sum_{i=1}^n y_i$ and $D^2 = n^{-1}\sum_{i=1}^n y_i^2$. This agrees with their Equation (39). The integral can be evaluated numerically, so posterior computations via Markov chain Monte Carlo, e.g., the Metropolis–Hastings algorithm, are relatively straightforward.

However, like in the satellite conjunction analysis example from the previous section, there is no guarantee that this posterior probability distribution for the interest parameter provides reliable inference. Here I will explore this question using a simulation study. Let Y consist of $n = 10$ iid samples from $\mathsf{N}(\mu, \sigma^2)$, with $\mu = 0.1$ and $\sigma = 1$; then the true value of ϕ is 10. Next, consider a hypothesis about ϕ, namely, $A = (-\infty, 9]$, which happens to be false. Let $\Pi_y(A)$ denote the probability of this hypothesis with respect to the posterior distribution in (2). Then $\Pi_Y(A)$ is a random variable, as a function of Y, and I am interested in its distribution function, $G_\theta(\alpha) = \mathsf{P}_{Y|\theta}\{\Pi_Y(A) \leq \alpha\}$. Figure 2 shows the estimated distribution function, based on 1000 data sets. The most striking feature of this plot is the high concentration of the distribution's mass around value 1. This is problematic because, remember, this is the posterior probability of a *false* hypothesis about ϕ. In those instances when $\Pi_y(A) \approx 1$ and, therefore, $\Pi_y(A^c) \approx 0$, the only reasonable conclusion would be that the data support hypothesis A. Since these problematic instances are apparently frequent, inferences drawn based on the posterior distribution described above would be systematically misleading.

It is worth mentioning that there is nothing special about the particular settings in this illustration. For any values of (n, μ, σ), there exists a hypothesis A, perhaps of similar form to the one above, such that, even if A is false, the posterior probability assigned to it will tend to be large; see Section 3.3.

Some statisticians might also recognize this normal coefficient of variation problem as one in the collection of challenging inference problems identified by Gleser and Hwang (1987) and Dufour (1997) and discussed in Bertanha and Moreira (2018). One could say that this and others in these classes are not a practical problems, and that might be true. But these "impossible" inference problems are precisely the ones we should use to test our foundations and intuition, in the same way that we test our understanding of measure theory by constructing and examining non-measurable sets, etc.

3.3   False confidence theorem

The previous two subsections showed that there are instances where inferences based on a Bayesian inferential model would be systematically misleading. It turns out that
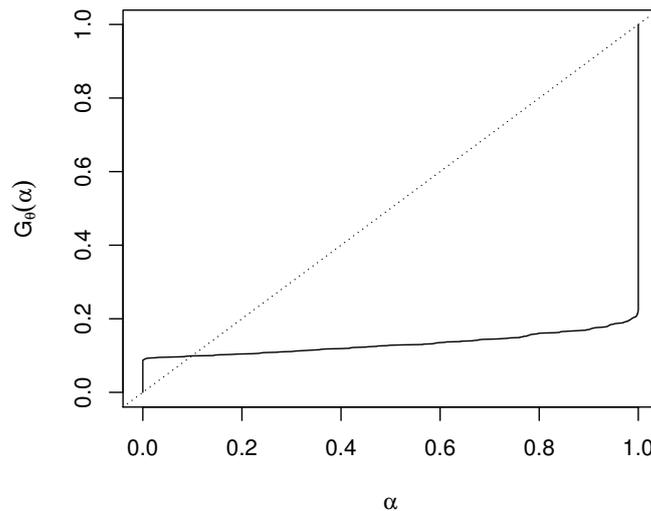
Figure 2: Plot of the estimated distribution function $G_\theta(\alpha) = \mathsf{P}_{Y|\theta}\{\Pi_Y(A) \leq \alpha\}$ of the Bayesian posterior probability, $\Pi_Y(A)$, where $A = (-\infty, 9]$ is a false hypothesis about the coefficient $\phi = \sigma/\mu$ when $\mu = 0.1$ and $\sigma = 1$.

these issues have nothing to do with the Bayesian features of those solutions. Instead, false confidence is a consequence of using *additive* degrees of beliefs, i.e., probabilities, to describe uncertainty. The following theorem makes this precise.

Theorem 1 (Balch et al. 2017). *Consider probability measure Π_y with a density function that is bounded and continuous for each y. Then for any $\theta \in \Theta$, any $\alpha \in (0,1)$, and any $p \in (0,1)$, there exists a set $A \subset \Theta$ such that*

$$A \not\ni \theta \quad and \quad \mathsf{P}_{Y|\theta}\{\Pi_Y(A) \geq 1 - \alpha\} \geq p.$$

Therefore, any inferential model with additive degrees of belief—including Bayes, fiducial, etc.—will be at risk of false confidence. That is, there are false hypotheses to which Π_y will tend to assign high probability, hence "systematically misleading conclusions." It is important to emphasize that the theorem only says *there exists* hypotheses afflicted by false confidence. Certainly, some of these hypotheses are trivial, e.g., complements of very small subsets of Θ, but the examples in Sections 3.1–3.2 reveal that non-trivial hypotheses can be affected too; see, also, Carmichael and Williams (2018). In my experience, the non-trivial yet problematic hypotheses have unusual shapes: the complement of a disc in the plane and the complement of a cone in the upper half-plane in Section 3.1 and 3.2, respectively. These unusual shapes most commonly arise when hypotheses concern a non-linear function of the full parameter. Similar concerns about the performance of Bayesian posterior inference for non-linear functions of θ were raised in Fraser (2011a) and Fraser et al. (2016), but Theorem 1 expands on their conclusions in several directions, in particular, it applies to any additive inferential model, not just Bayesian, and to more general hypotheses. Moreover, it helps to explain why, as advised by Cox (2006, p. 66) and others, it is dangerous to manipulate, e.g., marginalize, fiducial and confidence distributions according to the usual probability calculus.

When genuine prior information is available in the form of a proper prior distribution for θ, so that the behavior of $\Pi_Y(A)$ would be assessed with respect to the marginal

10

distribution of Y derived from the joint distribution of (Y, θ), then false confidence is also of no concern. For one thing, there are no *false hypotheses* in such a context since θ is not an unknown parameter in the marginal distribution of Y. Also, there are certain automatic controls on the distribution of $\Pi_Y(A)$ in such settings, but these bounds are generally only as meaningful as the postulated prior. See Section 5.4.

There is also a connection between the false confidence theorem and the non-existence of a truly "non-informative" prior distribution established in Proposition 2.1 of Martin and Liu (2016). That is, if every prior probability distribution is "informative" simply by virtue of its additivity, not because it's based on real prior information, then there will be false hypotheses towards which the posterior distribution is pushed.

In certain communities, additive probabilities are called *precise* in the sense that the postulated statistical model fully captures all the existing uncertainty.[6] Such an assumption is always difficult to justify but, for sure, said justification would require that the prior distribution be informative in some sense. With the commonly used default prior distributions, any claims of "precision" would be questionable:

> *[Bayes's formula] does not create real probabilities from hypothetical probabilities...* (Fraser 2014, p. 249)

There are so many hypotheses in the σ-algebra upon which the probability is defined and, given the necessarily limited informativeness of the data, it is unrealistic to expect that the posterior can assign meaningful or otherwise reliable probabilities to every one. Therefore, there must be hypotheses about which the posterior probabilities should not be trusted, and this is what the false confidence theorem says.

On its own, that there exists hypotheses afflicted by false confidence may not be a serious concern. The problem is that one generally does not know which hypotheses are afflicted and which are not. Arguably the strongest selling point of Bayesian and other approaches that produce probabilities as output is that they provide a means to answer *virtually any question* about the unknowns, specifically, that the full posterior can be marginalized to any quantity of interest to make justifiable inferences. The false confidence theorem says that this understanding is wrong, that is, there are features of θ for which certain questions cannot be answered reliably or justifiably; see, also, Grünwald (2018). Whether those afflicted hypotheses are of any practical interest depends on the context, but not knowing which ones they are puts the data analyst at risk. Short of identifying those afflicted hypotheses in each individual situation to mitigate risk, the only option is to avoid false confidence altogether, which, according to Theorem 1, requires breaking additivity. Not every non-additive belief will be free of false confidence, however, so some additional constraints are needed, as discussed in Section 4 below.

4 Validity condition

If false confidence is undesirable, as I say it is, then an important question is how to avoid it. Theorem 1 says additivity must be abandoned, and I will discuss specific types of non-additive degrees of belief in Section 6. But non-additivity alone is not enough to avoid false confidence so, before getting into those details, here I introduce a condition

---

[6]The *2019 International Symposium on Imprecise Probabilities*, http://isipta2019.ugent.be, has a theme, related to this point, that I like: "there's more to uncertainty than probabilities."

that will ensure false confidence can be avoided. The construction of an inferential model that achieves this condition is the topic of Section 7 below.

The inferential model formulation is one where inferences are drawn according to a data analyst's degree of belief based on data, a statistical model, etc. In a vacuum, I am free to construct and interpret my degrees of belief in any way that I please. But if the goal is to communicate my degrees of belief to others, to justify my conclusions, then some kind of objective or at least agreed-upon scale is needed in order to interpret degrees of belief. For example, if I conclude that hypothesis A is false because $b_y(A^c) > c$, then why might you find this convincing? It is not necessary that your degrees of belief agree 100% with mine, it is enough if I can demonstrate that the event $\{y : b_y(A^c) > c\}$ is rare if A is actually true. Of course, a more precise connection between the threshold "c" and the notion of "rare" are needed (see Definition 2) but the basic idea is familiar and uncontroversial. Indeed, Cox and Hinkley (1974, p. 45–46) write

> *We should not follow procedures which for some possible parameter values would give, in hypothetical repetitions, misleading conclusions most of the time.*

Recently, Reid and Cox (2015) echoed this point

> *Even if an empirical frequency-based view of probability is not used directly as a basis for inference, it is unacceptable if a procedure...of representing uncertain knowledge would, if used repeatedly, give systematically misleading conclusions.*

Both of these remarks mention "procedures" which seems to suggest a purely frequentist or behaviorist mode of reasoning based on rules and suggested actions, e.g., reject a hypothesis if and only if such and such happens. But Fisher (1973) and others (e.g., Mayo 1996, 2018; Mayo and Cox 2006) argue that there's more going on here, that the frequentist's error rate control leads to an inductive logic for justifying conclusions in individual instances. Indeed, Fisher (1973, p. 46) writes

> *...the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based ia fact communicable to, and verifiable by, other rational minds. The level of significance in such cases fulfils the conditions of a measure of the rational grounds for the disbelief it engenders.*

Along similar lines, Mayo (2018, p. 16) formalizes the notion of an argument from observation, through the absence of error, to conclusion as follows:

> *There is evidence an error is absent to the extent that a procedure with a very high capability of signaling the error, if and only if it is present, nevertheless detects no error.*

To avoid what Reid and Cox call "unacceptable," and to facilitate the inductive logic of Fisher and Mayo, some kind of calibration is needed, and I suggest the following constraint on ones inferential model.

Definition 2. An inferential model $(\mathscr{P}, y, \ldots) \mapsto b_y$ is *valid* if

$$\sup_{\theta \notin A} \mathsf{P}_{Y|\theta}\{b_Y(A) > 1 - \alpha\} \leq \alpha, \quad \forall \, \alpha \in [0,1], \quad \forall \, A \subset \Theta. \tag{3}$$

That is, if $\theta \notin A$, so that the hypothesis A is false, then $b_Y(A)$, as a function of $Y \sim \mathsf{P}_{Y|\theta}$, is stochastically no larger than $\mathsf{Unif}(0, 1)$.

The validity property (3) implies that, if A is false, then there is a small probability—with respect to the posited statistical model—that $b_Y(A)$ takes large values. In other words, data are unlikely to provide strong support for false assertions, therefore avoiding Reid and Cox's "systematically misleading conclusions." Since validity prevents the degrees of belief in a false hypothesis from being too large, in a distributional sense, it follows that a valid inferential model is not susceptible to false confidence. Similar calibration properties are discussed in Balch (2012) and Denœux and Li (2018).

Of course, since probabilities are susceptible to false confidence, they must automatically fail to be valid in the sense of Definition 2. However, this can be seen directly from the fact that (3) covers *all* assertions, including singletons and their complements. Indeed, if A is the complement of a singleton and b_y is a probability measure, then usually $b_y(A) = 1$ for all y, so (3) fails. But the examples in Section 3 demonstrate that it is not just these trivial hypotheses that can be afflicted by false confidence.

There are two specific questions that I would like to address. First, *why be concerned about all hypotheses?* Instead, one could ask that the inferential model satisfy a condition like that in (3) just for the one or few hypotheses of interest. There is no problem weakening the condition (3) to hold just for a specific collection of hypotheses, see Martin and Liu (2016, Def. 4.2), and indeed this would be satisfactory to the data analyst in a specific application. But at the meta level where data scientists are developing *methods* and corresponding software to used by others across applications, this strong control on performance is necessary. If I develop a method that automates the inferential model construction, with corresponding software that returns degrees of beliefs for any hypothesis, then unless I identify and warn the user about those problematic hypotheses, then I should assume that users will in fact use it for any hypothesis and protect myself from the associated risk. For example, if I recommend an inferential model that returns degrees of belief about $\theta = (\mu, \sigma)$ based on samples from $\mathsf{N}(\mu, \sigma^2)$, and I fail to warn the user about considering hypotheses of the form $A = \{(\mu, \sigma) : \sigma \leq c\mu\}$, for given $c \in \mathbb{R}$, then any erroneous conclusions based on this would be my responsibility. Consequently, the data scientist developing methods for mass consumption actually has *skin in the game* (Taleb 2018), potentially a lot, so in order to protect him/herself from the associated risk a strong, uniform control on performance like in (3), for any hypotheses the user might be interested in, would be desired.

Second, *why the specific "$1 - \alpha$" in (3)?* In other words, (3) could be written as

$$\sup_{\theta \notin A} \mathsf{P}_{Y|\theta}\{b_Y(A) > g_A(\alpha)\} \leq \alpha,$$

for some function $g_A(\alpha)$, so why not allow this more general notion of validity? Recall that a primary motivation for the validity condition was to have a scale on which the degrees of belief could be interpreted. To me at least, interpreting probabilities is a challenge because they are *relative*. For example, Nate Silver's *FiveThirtyEight* website estimated the (posterior) probability of a Donald Trump victory in the 2016 U.S. presidential election at 0.28, which was weak enough to predict that Hilary Clinton would win (e.g., Crane and Martin 2018e), whereas, in, say, the context of variable selection in regression, a (posterior) probability of 0.28 assigned to one of the 2^p possible subsets of the p predictor variables would be very strong support for that configuration. In the same way, the more general definition with $g_A(\alpha)$ in the above display would allow the scale of

interpretation to depend on the hypothesis A as well as on other specific features of the problem. To avoid the confusion of a moving scale of interpretation, an absolute scale is needed, but the specific choice of $\mathsf{Unif}(0,1)$ as an absolute scale is not essential—it's just familiar and natural.[7] See, also, Section 5.4.

An important consequence of the validity condition is that procedures designed based on the inferential model output will have frequentist guarantees. To see this clearly, define the dual to the inferential model's degrees of belief as

$$p_y(A) = 1 - b_y(A^c), \quad A \subseteq \Theta.$$

Of course, since the focus is on non-additive degrees of belief, p_y is different from b_y. The interpretation of p_y is as a measure of *plausibility*, which I discuss further in Section 6. Since (3) holds for all $A \subseteq \Theta$, there is an equivalent formulation in terms of p_y, i.e.,

$$\sup_{\theta \in A} \mathsf{P}_{Y|\theta}\{p_Y(A) \leq \alpha\} \leq \alpha, \quad \forall\, \alpha \in [0,1], \quad \forall\, A \subseteq \Theta. \tag{4}$$

Besides complementing the degrees of belief, p_y provides a simple and intuitive recipe to construct procedures with guaranteed frequentist error rate control.

Theorem 2. *Let $(\mathscr{P}, y, \ldots) \mapsto b_y$ be an inferential model that satisfies the validity condition in Definition 2. Fix $\alpha \in (0,1)$.*

(a) *Consider a testing problem with $H_0 : \theta \in A$ versus $H_1 : \theta \notin A$, where A is any subset of Θ. Then the test, T_y, that rejects H_0 if and only if $p_y(A) \leq \alpha$ has frequentist Type I error probability upper-bounded by α. That is,*

$$\sup_{\theta \in A} \mathsf{P}_{Y|\theta}(T_Y \text{ rejects } H_0) \leq \alpha.$$

(b) *Define the following data-dependent subset of Θ:*

$$C_\alpha(y) = \big\{\vartheta \in \Theta : p_y(\{\vartheta\}) > \alpha\big\}. \tag{5}$$

Then the frequentist coverage probability of $C_\alpha(Y)$ is lower-bounded by $1-\alpha$, making (5) a nominal $100(1-\alpha)\%$ confidence region. That is,

$$\inf_\theta \mathsf{P}_{Y|\theta}\{C_\alpha(Y) \ni \theta\} \geq 1 - \alpha.$$

This makes clear my previous claim that, if one has a valid inferential model, then decision procedures with good properties are immediate. It is worth pointing out again that Theorem 2(a) hold for all hypotheses A, even those problematic ones concerning a non-linear feature of the full parameter θ like the ones that cause the false confidence phenomenon in additive inferential models. Moreover, if $\phi = \phi(\theta)$ is a some feature of interest, then the set

$$\big\{\varphi \in \phi(\Theta) : \sup_{\vartheta:\phi(\vartheta)=\varphi} p_y(\{\vartheta\})\big\},$$

is nominal $100(1-\alpha)\%$ confidence region for ϕ. How to specifically construct a valid inferential model will be discussed in Section 7.

---

[7]To see what's special about the $\mathsf{Unif}(0,1)$ scale, consider the following example. Suppose you're visiting a place you've never been before and the weatherman on the TV says there's a 10% chance of rain. Likely you wouldn't be too concerned about not having an umbrella, but why? Maybe this weatherman reports *only* 0–10% chance of rain, no matter the circumstances, so that a 10% chance is actually very high. That this possibility seems totally absurd suggests that it's "natural" to think of probabilities as being calibrated to a $\mathsf{Unif}(0,1)$ scale, i.e., that it rains only 10% of the days the weatherman predicts a 10% chance of rain, and that's why I've imposed this in (3).

5 Remarks

5.1 BFFs and philosophy of science

Meng (2017) describes the efforts of a group working in and around the foundations of statistics and probability. That group goes by the name of "BFF" which has a double-meaning: the first, "Bayes, Frequentist, and Fiducial," is one that only a statistician might guess, while the second, "Best Friends Forever," is a modern pop culture term of endearment. The first two letters—B and F—are familiar to statisticians and below I briefly summarize these and make a connection to perspectives in philosophy of science.

- *Bayesian.* The defining feature of the Bayesian approach is that it is based strictly on probability calculus. That is, one starts with a statistical model that relates observable data to some unknown parameters, adds a prior probability distribution that is meant to describe uncertainty about those parameters, and then this prior is updated, via Bayes's formula, to a posterior distribution about the parameter given the observed data. The rationale behind the Bayesian approach to learning is that, roughly, with sufficiently informative data, perhaps after several iterations of this prior-to-posterior updating, those hypotheses which are "true" will have posterior probability close to 1, everything else having posterior probability close to 0. That is, learning is achieved by confirmation, in the spirit of Carnap (1962), Jeffreys (1961), and others, through the aggregation of evidence supporting a hypothesis.

- *Frequentist.* It is possible to talk about a Bayesian *approach* because the framework is normative, i.e., it gives a recipe for carrying out an analysis. It is much harder, however, to describe a frequentist *approach*, it is more of a philosophy or a sense in which any statistical approach or method can be evaluated. The idea is that, for a relevant scientific hypothesis, statistical methods should be good at assessing the plausibility of that hypothesis, where "good" is in the sense that the probability they detect departures between data and the hypothesis is low when there is none and high when there is. But since not detecting departures between data and a hypothesis does not imply truthfulness of the hypothesis, this strategy cannot be used for direct confirmation, only for refutation. That is, with sufficiently informative data and sufficiently good statistical methods, any hypothesis that is false is likely to be refuted. This lines up with the views of Popper (1959), Lakatos (1978), and Mayo (1996) who argue that knowledge is gained through falsification.

I have chosen to split the discussion of Bayesian and frequentist into two separate bullet points, but they are not mutually exclusive. Indeed, since frequentism is not an *approach*, but rather a means of evaluating a method, one can study the frequentist properties of Bayesian methods. It is often the case that Bayesian solutions agree with classical frequentist solutions asymptotically, e.g., a Bernstein–von Mises theorem holds, hence the Bayesian solutions also have the desirable frequentist properties asymptotically. In cases like these where Bayesian posterior probabilities have certain calibration properties, a verification–falsification balance can be achieved; see Box (1980) and Rubin (1984). However, these niceties are not universal since, as the normal coefficient of variation example, the false confidence theorem, and similar results in Fraser (2011a) and Fraser et al. (2016) indicate, these asymptotic guarantees may not be fully satisfactory and, as Gleser and Hwang (1987) remark, might even be misleading in certain cases.

Compared to the B and F above, it is more difficult to pin down precisely what the second F, *fiducial*, is. This is partly because, as Zabell (1992) elucidates, Fisher's

explanation of the "fiducial argument" changed significantly over time. At its genesis, Fisher's idea was to derive a sort of posterior distribution for the unknown parameter without a prior and without invoking Bayes's theorem, an ambitious goal. To see how this works, let θ be an unknown scalar parameter and suppose that $T = T(Y)$ is a scalar sufficient statistic, having distribution function F_θ. Take any $p \in [0, 1]$ and assume that the equation $p = F_\theta(t)$ can be solved uniquely for t, given θ, and for θ, given t. That is, assume there exists $t_p(\theta)$ and $\theta_p(t)$ such that

$$p = F_\theta(t_p(\theta)) = F_{\theta_p(t)}(t), \quad \forall\, (t, \theta).$$

If the sampling model is "monotone" in the sense that, for all (p, t, θ),

$$t_p(\theta) \geq t \iff \theta_p(t) \leq \theta,$$

then it is an easy calculation to show that

$$p = \mathsf{P}_{T|\theta}\{T \leq t_p(\theta)\} = \mathsf{P}_{T|\theta}\{\theta_p(T) \leq \theta\}. \tag{6}$$

The above expression is a mathematical fact, but has no direct connection to the inference problem because it doesn't depend on the observed $T = t$. Fisher's big idea was to fix $T = t$ and *define* the fiducial probability of the event $\{\theta \geq \theta_p(t)\}$ to be equal to p. In other words, for t the observed value of T, Fisher *defined*

$$\text{``}\mathsf{P}\{\theta \geq \theta_p(t)\}\text{''} = p, \quad \forall\, p \in [0, 1].$$

The collection $\{\theta_p(t) : p \in [0, 1]\}$, for fixed t, defines the quantiles of a distribution and, therefore, a distribution itself. Fisher called this the *fiducial distribution* of θ, given $T = t$. Naturally, since the fiducial probabilities are "derived" from calculations based on the sampling distribution, $\mathsf{P}_{T|\theta}$, of T for fixed θ, Fisher's interpretation of the fiducial probabilities was in a frequentist sense. So it's no surprise that the distinction between Fisher's fiducial quantile $\theta_p(t)$ and Neyman's $100p\%$ upper confidence limit for θ is blurry. Fisher ran into trouble, however, in the extension of his fiducial argument to multidimensional parameters. The challenges he faced were, in particular, non-uniqueness and a failure of the fiducial probability for some hypotheses to maintain the calibration property suggested by the connection to the sampling distribution in the expression (6). Notice that these challenges Fisher faced are both part of the high-level message coming out of the false confidence theorem in Section 3.3, i.e., that data is not informative enough to uniquely identify precise and reliable probabilities for every hypothesis. Fisher's response to these challenges was to back away from the sampling interpretation of his fiducial probabilities, opting instead for a more subjective, conditional view. This effort, however, was unsuccessful, which led to fiducial being labeled Fisher's "one great failure" (Zabell 1992, p. 369) and "biggest blunder" (Efron 1998, p. 105). Zabell (1992, p. 382) summarizes Fisher's motivation and efforts beautifully:

> *Fisher's attempt to steer a path between the Scylla of unconditional behaviorist methods which disavow any attempt at "inference" and the Charybdis of subjectivism in science was founded on important concerns, and his personal failure to arrive at a satisfactory solution to the problem means only that the problem remains unsolved, not that it does not exist.*

There are many other nice papers that discuss Fisher's views and insights on fiducial and inverse probability, among other things; see, e.g., Aldrich (2000, 2008), Edwards (1997), Savage (1976), Seidenfeld (1992), and Zabell (1989).

Both the aforementioned B and F correspond to well-established schools of thought in the philosophy of science, so what about fiducial? If the fiducial argument is difficult to understand even as a statistical approach or method, then understanding what, if any, philosophical position it corresponds to is all the more difficult. Maybe there's a new branch of philosophy of science waiting to be discovered. A fiducial-style philosophy of science would be similar to a confirmationist's view in the sense that it would be based on probabilities (or some other degree of belief measure) but also similar to a falsificationist's view in the sense that these probabilities can only be used as a tool for refutation, not for proof. In order to make such a theory work, I think some form of non-additivity would be necessary to handle the important and very common case where data *isn't inconsistent* with either A or A^c. I'll leave it to the experts in philosophy of science to decide if the precise formulation of such a theory is worth pursuing.

5.2   "The most important unresolved problem"

Even if it's not what Fisher had intended, at a high level, the fiducial argument can be viewed as an attempt to achieve some sort of middle-ground between the more widely-accepted B and F, one that would provide prior-free, data-dependent degrees of belief, like a Bayesian posterior, that simultaneously have certain frequentist calibration properties. Savage (1961) described the fiducial argument as "a bold attempt to make the Bayesian omelet without breaking the Bayesian eggs." As the quote from Zabell above communicates, the fact that Fisher's solution was unsuccessful doesn't mean that a satisfactory solution isn't possible or that the problem itself isn't interesting and worth pursuing. Indeed, Efron (2013b, p. 41) says that

> ...*perhaps the most important unresolved problem in statistical inference is the use of Bayes theorem in the absence of prior information.*

Similar comments can be found in Efron (1998) and, even more recently, during his featured address during the 2016 BFF workshop at Rutgers,[8] Efron referred to the above problem as the *Holy Grail* of statistics.

To be clear, there is no problem at all to get *a* posterior distribution without using a prior or Bayes's theorem. Any data-dependent probability distribution on the parameter space would be a candidate. What Efron is talking about is more than just getting a posterior distribution without a prior; he's asking for a "prior-free" procedure to construct a "posterior distribution" that is somehow reliable, that has performance guarantees. It's in this sense that my efforts here, which do not have anything directly to do with Bayes's theorem, are relevant to Efron's "most important unresolved problem."

My claim here is that the key to solving this most important problem is actually very simple: one just needs to be willing to abandon the requirement that degrees of belief be additive. As the false confidence theorem implies, however you construct your additive degrees of belief, there will exist hypotheses for which the probabilities lack the desired calibration and, consequently, would be "unacceptable" according to Reid and Cox. So, *going non-additive*, i.e., adopting a framework in which degrees of belief are measured

---

[8]https://statistics.rutgers.edu/bff2016

by some non-additive set function, is the only way to proceed towards a solution to this open problem. But non-additivity alone isn't enough. To avoid false confidence—and solve the fundamental problem—an additional constraint is needed, something like the validity condition in Section 4. There is, of course, no free lunch, so any new framework that can achieve this goal must necessarily make sacrifices elsewhere; see Section 9.

5.3   Beyond p and P

In Summer 2018, the statistics community was buzzing about p-values, their interpretation, and their limitations. Naturally, this got me thinking about the statistical "debates" pitting p-values and the NHST (null hypothesis significance testing) camp against posterior probabilities, Bayes factors, and the Bayesian camp. To set the scene, while p-values have received—and withstood—more than their fair share of scrutiny over the years (e.g., Berger and Sellke 1987; Schervish 1996), the discussion heated up when scientists started pointing fingers at the statistical community for low replication rates, etc. A bold move was made in 2015 when the editors of *Basic and Applied Social Psychology*, or *BASP*, banned the use of p-values—and much of the modern statistical toolbox—in their journal (Trafimow and Marks 2015). Naturally, this attracted media attention (e.g., Woolston 2015) and even sparked an unprecedented response from the American Statistical Association (Wasserstein and Lazar 2016), with a follow-up conference[9] and special issue of *The American Statistician*.[10]

My suspicion, which was confirmed by Professor David Trafimow in a personal communication, was that banning p-values was not a solution in and of itself but, rather, an impetus for change, forcing researchers to think harder about how they reason from data to conclusion, potentially leading to new ideas and higher quality research. Whether this strategy will be successful remains to be seen (e.g., Fricker Jr. et al. 2018), but almost anything is better than researchers "*p*-ing" everywhere (Valentine et al. 2015). Still, it makes sense to ask if contributions from the statistics community have been helpful. The dust is perhaps still settling, but, based on the expert commentary supplementing the ASA statement, the follow-up by Ionides et al. (2017), the recent proposal of Benjamin et al. (2017) to simply change the default "$p < 0.05$" cutoff to "$p < 0.005$," and the plethora of criticism that followed (e.g., Crane 2018b; Lakens et al. 2018; McShane et al. 2018; Trafimow et al. 2018), it seems that statisticians' views are as diverse as ever.

The response from the Bayesian or, at least, the non-NHST camp is that it's better in some sense to rely on probabilities for making inferences. As part of the justification for the p-value ban, Trafimow and Marks (2015) say

> the [p-value] fails to provide the probability of the null hypothesis, which is needed to provide a strong case for rejecting it.

That a p-value is *not* the probability that the null hypothesis is true is, of course, a mathematical fact. But is it really necessary to have such a probability for making inferences? Fisher says no:

> As Bayes perceived, the concept of mathematical probability affords a means, in some cases, of expressing inferences from observational data, involving a degree of uncertainty, and of expressing them rigorously... yet it is by no

---

[9]https://ww2.amstat.org/meetings/ssi/2017/
[10]https://ww2.amstat.org/meetings/ssi/2017/callforpapers.cfm

> *means axiomatic that the appropriate inferences... should always be rigorously*
> *expressible in terms of this same concept.* (Fisher 1973, p. 40)

And about the p-value—or observed significance level—he says

> *It is more primitive, or elemental than, and does not justify, any exact prob-*
> *ability statement about the proposition.* (Fisher 1973, p. 46)

But before asking about whether probability is a necessity, an even more basic question is if a genuine probability is available in general. If we put together the famously provocative quote from de Finetti (1990)—*probability does not exist*—with the refreshingly blunt remark of Fraser's on page 11 above, then we start to see how concerning this is.[11] If probabilities exist only in our minds, and don't become "real" simply by updating via Bayes's formula, then how can they be justified as relevant to a scientific problem? For example, reconsider the highly-publicized predictions of 2016 U.S. presidential election by Nate Silver on his *FiveThirtyEight* website/blog (e.g., Crane and Martin 2018e). On the day of the election, Silver predicted Hilary Clinton would win with probability 0.72 and, therefore, Donald Trump would win with probability 0.28. These numbers are based on specified prior distributions, statistical models, and polling data, all combined via Bayes's formula and evaluated on a computer running Markov chain Monte Carlo. Does Silver making such a probability statement suddenly imply that the election will be determined by the flip of a 72/28 coin? Of course not. The numbers may be meaningful to Silver himself in many ways, e.g., prices for bets he's willing to make, but the relevance of Silver's numbers to me, to any other consumer, or to "reality" must be established. As I see it, there are only two ways this relevance argument can be made: either I understand and agree with all of Silver's modeling assumptions, and accept his subjective probabilities as my own, or I simply trust that Silver's predictions are accurate or otherwise reliable. The former is quite unlikely, especially for consumers without statistical training; the latter is a choice that I'm free to make or not, though recent results suggest that Silver's predictions may not be so reliable (Crane 2018a,c). In any case, (posterior) probabilities are not objective or otherwise obviously relevant to the scientific problem at hand, their relevance has to be established in some way.

While I am skeptical about being able to justify the relevance of a probability distribution to a scientific problem,[12] it is worth pointing out that there would be a tremendous payoff if, in fact, it could be done. Arguably, the more natural way to think about knowledge is as a process by which evidence piles up in favor of some hypothesis until which its truthfulness is undisputed. This confirmationinst style of learning is quite powerful because one effectively has access to "truth" via probability ≈ 0 or ≈ 1 events and what Shafer (2016) calls *Cournot's Principle*; see also Shafer (2007) and Shafer and Vovk (2006). Popper's falsificationist view emerged, not because it's more powerful, but because, like me, he doubted the confirmationist's premise that a meaningful probability distribution—one that could be continuously updated via Bayes's theorem, eventually pointing to "truth"—generally exists. Popper's only option, therefore, is to discard the

---

[11]To be fair, Trafimow and Marks (2015) acknowledge that the use of posterior distributions based on default priors are not satisfactory substitutes for genuine prior probabilities.

[12]Consider an ideal case where a *true* and, therefore, relevant prior exists, e.g., with exchangeable sequences where de Finetti's theorem applies (de Finetti 1937; Hewitt and Savage 1955). Even in such cases, the true prior, whose existence is guaranteed by the theorem, depends on the data, so one actually needs an infinite—or at least long—data sequence to identify the true prior.

strong existence assumption, but the price he pays is not being able to judge a hypothesis as "true," only as "not yet proven false."

Back to Summer 2018. Given the "p versus P" debates going on at the time, and the fact that this focus on NHST versus Bayes seemed to be misaligned, I planned to write a non-technical paper about moving forward, "beyond p and P," by thinking in terms of a third "p"—plausibility—that, when equipped with the validity condition, can achieve much of what both the frequentists and Bayesians want. Not long after my plan for this crystallized, Deborah Mayo's book (Mayo 2018) appeared wherein she talks about getting "beyond the Stat Wars." Aside from having a more clever name—*Stat Wars*—for the controversy, her presentation is much more comprehensive than I could do, so I decided to take my efforts in a different direction, hence this manuscript.

5.4   Bayesian validity?

According to the false confidence theorem, validity is not compatible with a Bayesian inferential model or, for that matter, any other the produces additive degrees of belief as output. But the validity condition, in Equation (3), is rather strong, so one could ask if the Bayesian approach satisfies perhaps some weaker form of the validity condition. The answer to this question depends on what kind of Bayesian one is.

As I see it, there are two kinds of Bayesians: those who believe in the prior and those who don't. By "believe in the prior" I mean that one would actually bet based on their prior probabilities, would effectively dismiss hypotheses with low prior probability, and/or would be satisfied in evaluating the performance of their posterior distribution with respect to the corresponding marginal distribution of Y, i.e., $\mathsf{P}_Y = \int \mathsf{P}_{Y|\theta}\,\Pi(d\theta)$, where Π is the prior distribution for θ. If one believes in the prior in this sense, then a *Bayesian validity property* comes for free. That is, by Markov's inequality and the martingale property of the posterior with respect to the P_Y marginal,

$$\mathsf{P}_Y\{\Pi_Y(A) > \alpha^{-1}\Pi(A)\} \leq \alpha, \quad \text{for any } \alpha \in (0,1) \text{ and any } A. \tag{7}$$

In words, the posterior probability being a large multiple of the prior probability is a rare event with respect to P_Y. Is there anything more than a superficial similarity between the inequalities in (7) and that in (3)? While there's no notion of a hypothesis about θ being "true" or "false" in this context, the closest thing is to consider those hypotheses with small prior probability, since a believer in their prior probabilities would consider these to be "effectively false." Then (7) says that it's rare to see data that provide strong support for hypotheses having small prior probability, which is a version of the validity property. These self-consistency properties are expected, that is, if the prior is meaningful, then the posterior obtained by the usual probability calculus is equally meaningful.

The second group—those who don't believe in the prior—consists of those Bayesians who might, for simplicity, use default or non-informative priors, trusting that the prior-to-posterior updates will correct for any shortcomings in the prior specification. In order for the marginal distribution and the prior probabilities in (7) all make sense, let me assume that the prior Π is proper but diffuse, as would often be the case in situations where genuine prior information is lacking. It's possible that P_Y could be very different from the true distribution of Y, in which case (7) isn't meaningful at all. But let me also assume that P_Y is a reasonable approximation of the true distribution, e.g., it passes certain prior- or posterior-predictive checks (e.g., Box 1980; Gelman et al. 1996; Rubin 1984). Even in this case, there is still trouble coming from (7). The inequality says that

the posterior probabilities are unlikely to be much larger than the prior probabilities. But virtually any "non-extreme" hypothesis A has small $\Pi(A)$ by virtue of the diffuse non-informative prior, *not because there's reason to believe that it's false.* So if a potentially true hypothesis has small prior probability, and it's unlikely that the posterior will be much larger than the prior, then how can one expect to learn? This is precisely what's going on in the satellite conjunction analysis example summarized in Section 3.1. The diffuse prior behind the scenes assigns small prior probability to the collision event and, even if collision is imminent, (7) implies that the posterior probabilities are unlikely to turn large, hence false confidence. So, the inequality (7) that is an asset to the Bayesian who believes in his/her prior becomes a liability to the Bayesian who doesn't.

6 Non-additive beliefs

6.1 Lower probabilities and capacities

In the previous sections I have argued that, at least in order to avoid false confidence, it is necessary that the inferential model output be *non-additive*, e.g., that $b_y(A) + b_y(A^c) \neq 1$ for some hypothesis A. This section describe some particular non-additive set functions that have been suggested in the literature as measures of degrees of belief, all of which generalize the familiar additive probability. (Here it is not necessary to consider beliefs that depend on data y so I will write b instead of b_y throughout.)

The most general non-additive measure considered here is a *lower probability*. Let Θ be completely metrizable and consider a collection $\mathscr{M}$ of probability distributions on the measurable space $(\Theta, \mathcal{B}_\Theta)$. Now define the lower probability as

$$b(A) = \inf_{P \in \mathscr{M}} P(A), \quad A \in \mathcal{B}_\Theta.$$

Of course, there is a dual to b, namely,

$$p(A) = \sup_{P \in \mathscr{M}} P(A) = 1 - b(A^c),$$

which in the this context is called the corresponding *upper probability*. The lower probability is non-additive in the sense above or, more specifically, they are *super-additive* in the sense that

$$b(A \cup B) \geq b(A) + b(B) - b(A \cap B),$$

with strict inequality at least for some A and B. Super-additivity implies that $b(A) \leq p(A)$, and this gap between the lower and upper probabilities can be understood as a measure of how imprecise the imprecise model is; Dempster (2008) referred to the difference $p(A) - b(A)$ as the "don't know" probability. Walley (1991) presents a theory of imprecise probability based on lower probabilities—or, more generally, *lower provisions*—with a focus on achieving a *coherence*, i.e., that a gambler who determines the prices she's willing to pay to make certain bets according to b can't be made a sure loser. In fact, if $\mathscr{M}$ is closed and convex, then the lower probability b is coherent (e.g., Walley 2000).

A special case of lower probabilities that often arises in statistical applications, namely, in robustness studies, are *Choquet capacities* or, *capacities* for short (Choquet 1954). Let b be a set function defined on a domain $\mathcal{A} \subseteq 2^\Theta$, with the basic properties $b(\varnothing) = 0$ and $b(\Theta)$. Then b is a *capacity of order n* if

$$b(A) \geq \sum_{\varnothing \neq I \subseteq \{1,\ldots,n\}} (-1)^{|I|-1} b\Big(\bigcap_{i \in I} A_i\Big)$$

for every $A \in \mathcal{A}$ and every collection $\{A_1, \ldots, A_n\}$ of sets in $\mathcal{A}$ with $A_i \subset A$. Such a function b is said to be *n-monotone*. Then the corresponding dual, $p(A) = 1 - b(A^c)$, is *n-alternating* and satisfies

$$p(A) \leq \sum_{\varnothing \neq I \subseteq \{1, \ldots, n\}} (-1)^{|I|-1} p\left(\bigcup_{i \in I} A_i\right).$$

Since capacities are special cases of lower probabilities/provisions, it follows that they too are super-additive. It is also clear that every capacity of order n is also a capacity of order $n - 1$ so the class is shrinking in n. Therefore, capacities of order 2 are the most general and have been investigated by statisticians in, e.g., Huber and Strassen (1973), Huber (1973, 1981), Wasserman (1990), and Wasserman and Kadane (1990).

A relevant question is if there are any obvious practical advantages to super-additivity? The answer is yes, and the simplest is that of modeling *ignorance*. Despite being somewhat extreme, ignorance is apparently a practically relevant case since the developments of default priors are motivated by examples where the data analyst is ignorant about the parameter, i.e., no genuine prior information is available. So, understanding how to describe ignorance mathematically ought to be insightful for understanding the advantages and limitations of prior-free probabilistic inference. A first question is: *what does ignorance mean?* A standard assumption is that the universe Θ is given—hence not fully ignorant—and contains at least two points. Subject to these conditions, ignorance means that there is no evidence available that supports the truthfulness or falsity of any non-trivial hypothesis, and a natural way to encode this mathematically is by setting $b(A) = 0$ for any proper subset A of Θ, and $b(\Theta) = 1$. Shafer (1976) and others call this a *vacuous belief*, and it's easy to check that this b is super-additive. Compare this to standard approaches to formulating ignorance in classical probability, based on assumptions of symmetry/invariance or Laplace's "principle of indifference," both of which are based on certain judgments—or *knowledge*—and, therefore, don't really model ignorance.

When working with ordinary probabilities, updating via conditioning is straightforward and uncontroversial. But the analogous updating of lower probabilities, capacities, and even the belief functions discussed in Section 6.2 can have some arguably unexpected behaviors, e.g., *dilation*. See Gong and Meng (2017) for a modern discussion of these updating rules and their respective properties. Some comments on Dempster's rule of combination (see, e.g., Shafer 1976) in the context of combining inferential models represented as belief functions will be given in Section 8.2 below.

6.2   Belief functions

A *belief function* is a special type of capacity, one that is ∞-monotone, i.e., *n*-monotone for all $n \geq 1$. This is a strong condition, but there are some advantages that come with this specialization, namely, that capacities are quite complicated but belief functions can be characterized in terms of simpler and more intuitive mathematical objects.

As is customary in the literature on belief functions, I'll assume for the moment that Θ is a finite space. Then the power set 2^Θ is also finite and I can imagine a probability mass function m defined on 2^Θ, assumed to satisfy $m(\varnothing) = 0$. For $A \subseteq \Theta$, the quantity $m(A)$ encodes the degree of belief in the truthfulness of A but in no proper subset of A. Then the function

$$b(A) = \sum_{B : B \subseteq A} m(B), \quad A \subseteq \Theta, \tag{8}$$

is a belief function, i.e., is a capacity of order ∞. Intuitively, the belief in the truthfulness of the assertion A is defined as the totality of the degrees of belief in A and all its proper subsets. It turns out that there is a one-to-one correspondence between the mass function m and the belief function b; see Shafer (1976) or Yager and Liu (2008) for details.

A further insight that can be gleaned from (8) is that the belief function is effectively characterized by a *random set*. That is, one can think of m as the mass function for a random element taking values in 2^Θ. But this equivalence between belief functions and random sets only holds up in the case when Θ is a finite set. When Θ is infinite, it turns out that random sets corresponding to a special class of belief functions, which I describe in the next subsection. To make the connection between random sets and belief functions in the general case, there is an extra layer, what Shafer (1979) calls an *allocation of probability*, which I will not discuss here. Kohlas and Monney (1995) also discuss this gap between general belief functions and those based on random sets.

My discussion of belief functions here does nowhere near justice to the tremendous amount of incredible work that has been done in this area. And since I can't possibly provide an adequate overview of the literature, let me just point the reader to the 2016 special issue[13] of the *International Journal of Approximate Reasoning*, edited by Thierry Denœux, to celebrate 40 years since publication of Glenn Shafer's monograph, which gives a look back at the belief function developments over that time.

6.3 Random sets

As eluded to in the previous subsection, random sets lead to belief functions but not all belief functions correspond to random sets; the equivalence only holds when the domain Θ is a finite set. In general, a random set on an infinite domain will inherit certain continuity properties that a belief function need not satisfy.

As expected, defining random sets in a mathematically rigorous way requires considerable care. Since this amount of rigor isn't necessary—or helpful—for my present purposes, I will not be completely precise in what follows. The reader interested in the precise details can check out Nguyen (1978, 2006) or Molchanov (2005). Let $\mathcal{T}$ denote a random set in Θ, i.e., a (measurable) function from a sample space to 2^Θ. If $\mathsf{P}_\mathcal{T}$ is the distribution of $\mathcal{T}$, define the *containment functional*

$$c(A) = \mathsf{P}_\mathcal{T}(\mathcal{T} \subseteq A), \quad A \subseteq \Theta.$$

This containment functional obviously satisfies $c(\varnothing) = 0$ and $c(\Theta) = 1$. It can also be shown that c is ∞-monotone, so c is a belief function. However, through its connection to a genuine probability measure, c is also continuous in the sense that

$$c(A_n) \downarrow c(A), \quad \text{for any } A_n \downarrow A,$$

a condition which is not required of general belief functions. Interestingly, the additional layer that Shafer needed to connect belief functions to random sets is entirely due to the latter being continuous. Indeed, according to the famous *Choquet capacity theorem*, every continuous belief is the capacity functional of a random set. In Section 7, I will describe the construction of a valid inferential model whose output is a random set containment function, hence is equivalent to a framework built around continuous belief functions.

---

[13]https://www.sciencedirect.com/journal/international-journal-of-approximate-reasoning/special-issue/10BG01ZSM7P

An even more specific class of belief functions is that based on random sets, $\mathcal{T}$ which are *nested* in the sense that, for any pair of sets T and T' in the support of $\mathcal{T}$, either $T \subseteq T'$ or $T \supseteq T'$. Then the corresponding belief function, b, is called *consonant*, and the dual, p, to b has the following surprising property:

$$p(A) = \sup_{\vartheta \in \Theta} p(\{\vartheta\}).$$

That is, the plausibility and, hence, the belief function are determined by an ordinary point function, called the *plausibility contour*,

$$p(\{\vartheta\}) = \mathsf{P}_{\mathcal{T}}(\mathcal{T} \ni \vartheta), \quad \vartheta \in \Theta.$$

The belief/plausibility functions corresponding to nested random sets are also known as *necessity/possibility measures* (e.g., Dubois and Prade 1988).

7 Valid inferential models

7.1 Basic construction

The previous sections have focused primarily on the point that non-additive degrees of belief satisfying a validity constraint (or something similar) are needed in order to avoid false confidence. Obvious questions that come to mind are *does a valid inferential model exist?* and, if so, *how to construct it?* This section reviews the detailed construction first presented in Martin and Liu (2013) and then expanded upon in Martin and Liu (2016).

The inferential model depends explicitly on the posited statistical model, so the starting point for the following construction is a particular representation of that model, as a data-generating process. That is, write

$$Y = a(\theta, U), \quad U \sim \mathsf{P}_U, \tag{9}$$

where $U \in \mathbb{U}$ is called an *auxiliary variable* with distribution P_U fully known, and $a : \Theta \times \mathbb{U} \to \mathbb{Y}$ is a known mapping. The representation in (9) is called an *association* between data Y, parameter θ, and auxiliary variable U. An association always exists since any sampling model that can be simulated on a computer must have a representation (9), but that representation is not unique. The association need not be based on knowledge of the actual data-generating process, it can be viewed simply as a representation of the data analyst's uncertainty; that is, no claim that U exists in "real life" is made or required. The advantage to introducing U in (9) is that it identifies a fixed probability space, namely, $(\mathbb{U}, \mathscr{B}, \mathsf{P}_U)$, for a σ-algebra $\mathscr{B}$, on which we are free to carry out probability calculations independent of the data and parameter; as I describe below, these "probability calculations" are with respect to the distribution of a random set on $\mathbb{U}$.

Readers familiar with Fisher's fiducial argument, or the structural inference framework of Fraser (1968), will recognize the auxiliary variable U as a "pivotal quantity." Roughly speaking, these approaches aim to solve the equation (9) for θ and then use the distribution $U \sim \mathsf{P}_U$ and the observed value of Y to induce a distribution for θ. This fiducial argument will lead to an additive probability distribution on Θ which does not satisfy the validity property. Even the extended fiducial argument put forth by Dempster, where the inferential output need not be additive, the validity property generally fails.

Therefore, something extra is needed. More discussion on the connection between the fiducial argument and the approach presented here is given in Section 8.1.

According to the formulation in (9), there is a quantity U that, together with θ, determines the value of Y. If both Y and U were observable, then (9) could be simply solved for θ, hence the best possible inference. That is, if $Y = y$ and $U = u$ were observed, then it would be *certain* that the true θ belongs to the set

$$\Theta_y(u) = \{\vartheta \in \Theta : y = a(\vartheta, u)\}.$$

However, since U is not observable, the best one can do is to accurately *predict*[14] the unobserved value using some information about its distribution P_U.

To accurately predict the specific unobserved value of U, it is not enough to use an independent draw from P_U as the fiducial argument suggests. Intuitively, hitting a target with positive probability requires stretching that draw from P_U into a genuine set. Towards formalizing this notion of "stretching a draw from P_U," consider a random set, $\mathcal{S}$, having distribution $\mathsf{P}_{\mathcal{S}}$, with the property that it contains a large P_U-proportion of U values with high probability. More precisely, define $\gamma_{\mathcal{S}}(u) = \mathsf{P}_{\mathcal{S}}(\mathcal{S} \ni u)$, a feature of the distribution $\mathsf{P}_{\mathcal{S}}$, called the contour function of $\mathcal{S}$, and suppose that

$$\gamma_{\mathcal{S}}(U) \text{ is stochastically no smaller than } \mathsf{Unif}(0,1). \tag{10}$$

In other words, $\mathcal{S}$—or, rather, P_S—must be such that

$$\mathsf{P}_U\{\gamma_{\mathcal{S}}(U) \leq \alpha\} \leq \alpha, \quad \forall\, \alpha \in [0,1].$$

Intuitively, this condition suggests that $\mathcal{S}$ can cover a draw U from P_U with high probability. Martin and Liu (2013) demonstrate that this is a relatively mild condition and provide some general strategies for constructing such a set $\mathcal{S}$. For example, if P_U is non-atomic and $h : \mathbb{U} \to \mathbb{R}$ is a continuous function, non-constant on any set with positive P_U-measure, then $h(U)$ is a continuous random variable and the random set

$$\mathcal{S} := \{u \in \mathbb{U} : h(u) \leq h(\tilde{U})\}, \quad \tilde{U} \sim \mathsf{P}_U, \tag{11}$$

satisfies (10); see Martin and Liu (2016, p. 61) for a proof. One case that arises in a number of examples is where $\mathbb{U} = [0,1]$ and $\mathsf{P}_U = \mathsf{Unif}(0,1)$, and a "default" choice of $\mathcal{S}$ is like that given above, with $h(u) = |u - \frac{1}{2}|$, that is,

$$\mathcal{S} = \{u \in (0,1) : |u - 0.5| \leq |U - 0.5|\}, \quad U \sim \mathsf{Unif}(0,1). \tag{12}$$

Now, if $\mathcal{S}$ provides a reliable guess of where the unobserved U resides, in the sense of (10), then an equally reliable guess of where the unknown θ resides, given $Y = y$, is

$$\Theta_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_y(u),$$

another random set, with distribution determined by $\mathsf{P}_{\mathcal{S}}$. It is, therefore, natural to say that data y supports the claim "$\theta \in A$" if $\Theta_y(\mathcal{S})$ is a subset of A. Then the inferential model output is

$$b_y(A) = \mathsf{P}_{\mathcal{S}}\{\Theta_y(\mathcal{S}) \subseteq A\}, \quad A \subseteq \Theta. \tag{13}$$

---

[14]Perhaps *guess* is a better word to describe this operation than *predict* but the former has a particularly non-scientific connotation, which I'd like to avoid.

This is a belief function, like in Section 6.2, so it inherits those properties, e.g., it's super-additive and has a corresponding plausibility function $p_y(A) = 1 - b_y(A^c)$.

It will typically be the case that $\Theta_y(\mathcal{S})$ is *nested*, as discussed in Section 6.3; I'll talk about why in Section 8.2. Consequently, the belief function b_y is consonant and fully determined by the plausibility contour $p_y(\{\vartheta\}) = \mathsf{P}_\mathcal{S}\{\Theta_y(\mathcal{S}) \ni \vartheta\}$, $\vartheta \in \Theta$. According to Shafer (1976, Ch. 10), while the use of a consonant belief function as a description of evidence might be questionable in general, one case where it would be appropriate is in describing inferential evidence, as is the focus here.

Next are a couple simple examples to illustrate the construction and the corresponding belief/plausibility function evaluation.

Example 1. Let Y be a scalar with continuous distribution function F_θ, depending on a scalar parameter θ. Then a natural choice for the association (9) would be $Y = F_\theta^{-1}(U)$, where $U \sim \mathsf{Unif}(0, 1)$. Then $\Theta_y(u) = \{\vartheta : F_\vartheta(y) = u\}$ and, with a random set $\mathcal{S}$, it is easy to check that $\Theta_y(\mathcal{S}) = \{\vartheta : \mathcal{S} \ni F_\vartheta(y)\}$, so

$$p_y(\{\vartheta\}) = \mathsf{P}_\mathcal{S}\{\mathcal{S} \ni F_\vartheta(y)\}.$$

Using the default random set $\mathcal{S}$ in (12), this can be simplified:

$$p_y(\{\vartheta\}) = 1 - |2F_\vartheta(y) - 1|.$$

From here, plausibility of any hypothesis A can be evaluated as

$$\mathsf{pl}_y(A) = \sup_{\vartheta \in A} \mathsf{pl}_y(\{\vartheta\}),$$

and hence belief can be evaluated according to the formula $b_y(A) = 1 - p_y(A^c)$. For example, suppose that F_θ is the distribution function of $\mathsf{Gamma}(\theta, 1)$, where $\theta > 0$ is the unknown shape parameter. Panel (a) in Figure 3 shows the plausibility contour and Panel (b) shows $b_y(A_\vartheta)$ and $p_y(A_\vartheta)$ where $A_\vartheta = (0, \vartheta)$, sort of lower and upper distribution functions. Note the gap between the lower and upper in Panel (b).

Example 2. Consider a binomial model, $Y \sim \mathsf{Bin}(n, \theta)$, where n is a known positive integer and $\theta \in (0, 1)$ is the unknown success probability. Despite its apparent simplicity, inference on θ in the binomial model is far from trivial (e.g., Brown et al. 2001). In this case, the most natural version of (9) is

$$F_\theta(Y - 1) \leq U < F_\theta(Y), \quad U \sim \mathsf{Unif}(0, 1),$$

where F_θ is the $\mathsf{Bin}(n, \theta)$ distribution function. There is no simple equation linking (Y, θ, U) in this case, just a rule "$Y = a(\theta, U)$" for producing Y with given θ and U. This determines the map $u \mapsto \Theta_y(u)$, which is given by

$$\begin{aligned}
\Theta_y(u) &= \{\vartheta : F_\vartheta(y - 1) \leq u < F_\vartheta(y)\} \\
&= \{\vartheta : G_{n-y+1,y}(1 - \vartheta) \leq u < G_{n-y,y+1}(1 - \vartheta)\} \\
&= \{\vartheta : 1 - G_{n-y+1,y}^{-1}(u) \leq \vartheta < 1 - G_{n-y,y+1}^{-1}(u)\},
\end{aligned}$$

where $G_{a,b}$ denotes the $\mathsf{Beta}(a, b)$ distribution function. This calculation follows from the connection between the binomial and beta distribution functions, namely, $F_\theta(y) = G_{n-y,y+1}(1 - \theta)$, which can be derived using integration-by-parts. Note that, in this

(a) Plausibility contour

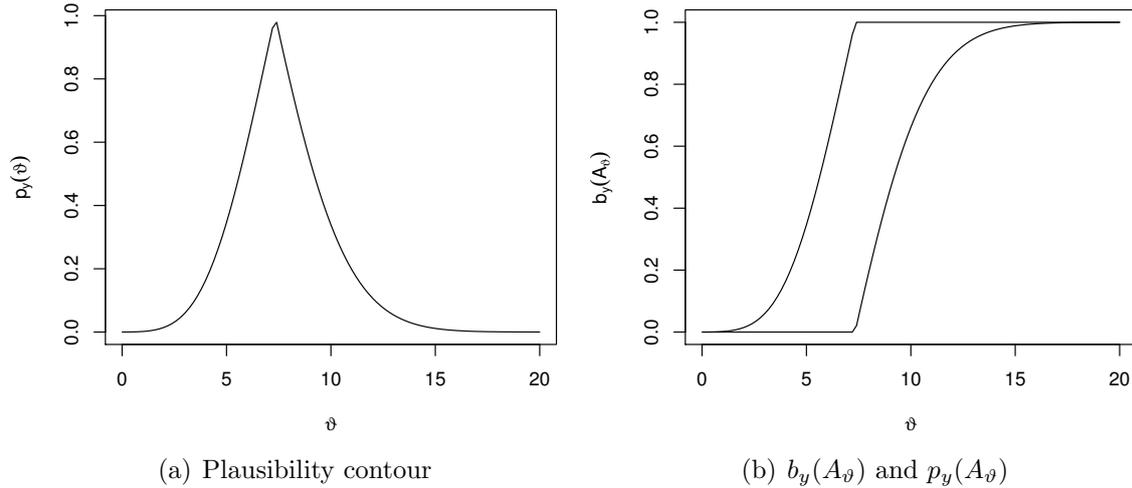(b) $b_y(A_\vartheta)$ and $p_y(A_\vartheta)$

Figure 3: Results of the gamma inference problem in Example 1 with $Y = 7$. Panel (a) shows the plausibility contour and Panel (b) shows the belief and plausibility functions for the collection of hypotheses $A_\vartheta = (0, \vartheta)$.

discrete-data setting, the set $\Theta_y(u)$ is an interval, not a singleton like in the continuous-data example above. Next, for simplicity, again I'll use the default random set $\mathcal{S}$ in (12) for predicting the unobserved $U \sim \mathsf{Unif}(0, 1)$. Combining $\mathcal{S}$ with $\Theta_y(\cdot)$ above gives

$$\Theta_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_y(u) = \left[1 - G_{n-y+1,y}^{-1}(\tfrac{1}{2} + |\tilde{U} - \tfrac{1}{2}|), 1 - G_{n-y,y+1}^{-1}(\tfrac{1}{2} - |\tilde{U} - \tfrac{1}{2}|)\right],$$

where $\tilde{U} \sim \mathsf{Unif}(0, 1)$. Then the probability, $\mathsf{P}_\mathcal{S}\{\Theta_y(\mathcal{S}) \subset A\}$ can be computed for any A using the distribution of the endpoints of $\Theta_y(\mathcal{S})$, which are simple functions of a uniform random variable. For example, the plausibility contour equals

$$p_y(\{\vartheta\}) = \mathsf{P}_\mathcal{S}\{\Theta_y(\mathcal{S}) \ni \vartheta\} = 1 - \mathsf{P}_\mathcal{S}\{\max \Theta_y(\mathcal{S}) < \vartheta\} - \mathsf{P}_\mathcal{S}\{\min \Theta_y(\mathcal{S}) > \vartheta\}.$$

The expression for this is messy, but the computation is easy. Figure 4(a) shows a plot of this plausibility contour based on and observed $y = 7$ and $n = 18$. A feature that distinguishes this plausibility contour from that in Example 1 is the plateau at the top; this is a consequence of $\Theta_y(u)$ being an interval for all u, which itself is a consequence of the discreteness of Y. Panel (b) shows the belief and plausibility for a sequence of one-sided hypotheses like that in Figure 3(b).

Importantly, the inferential model obtained by the above construction, driven by the random set $\mathcal{S} \sim \mathsf{P}_\mathcal{S}$ that satisfies (10), is provably valid in the sense of Definition 2.

Theorem 3 (Martin and Liu 2013). *Suppose that $\mathcal{S}$ satisfies* (10) *and that $\Theta_y(\mathcal{S})$ is non-empty with $\mathsf{P}_\mathcal{S}$-probability 1 for each y. Then the inferential model with output b_y defined in* (13) *is valid in the sense of Definition 2.*

Recall that the validity condition was designed to properly calibrate the data analyst's beliefs for scientific application, to avoid false confidence. An important practical consequence, presented in Theorem 2 above, is that it's straightforward to construct decision procedures with guaranteed frequentist error rate control. For example, the corresponding plausibility region (5) attains the nominal frequentist coverage probability.

27

(a) Plausibility contour
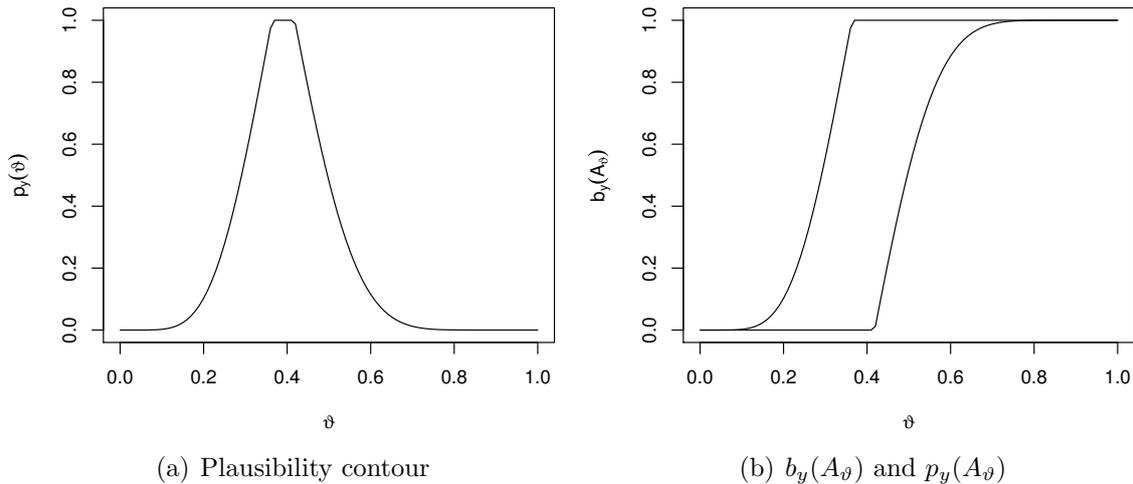
(b) $b_y(A_\vartheta)$ and $p_y(A_\vartheta)$

Figure 4: Results of the binomial inference problem in Example 2 with $y = 7$ and $n = 18$. Panel (a) shows the plausibility contour and Panel (b) shows the belief and plausibility functions for the collection of hypotheses $A_\vartheta = (0, \vartheta)$.

The non-emptiness assumption in the theorem often holds automatically for natural choices of $\mathcal{S}$, but can fail if the dimensions of θ and U don't match up (see Section 7.2) or in problems that involve non-trivial parameter constraints. When $\mathsf{P}_\mathcal{S}\{\Theta_x(\mathcal{S}) \neq \varnothing\} < 1$, Ermini Leaf and Liu (2012) propose to use an *elastic* version of $\mathcal{S}$ that stretches just enough that $\Theta_y(\mathcal{S})$ is non-empty, and they prove a corresponding validity theorem.

To conclude this brief introduction, I should mention that the approach described above is not necessarily the only way to construct a valid inferential model. I'm currently not aware of any competing constructions, but I make no claims that what I've presented above is the best or only way to achieve the desired validity property.

7.2 Dimension reduction

The basic construction presented above is actually quite general, but depending on the data structure and/or the objectives of the analysis, it may be possible to make the inferential model more efficient by reducing the dimension of the auxiliary variable that links the data and parameter in (9). This dimension reduction strategy has proved to be useful in two frequently encountered scenarios:

- when data consists of individual components each contributing some information about a common parameter, e.g., like with iid data, and

- when the parameter of interest is some lower-dimensional feature, $\phi = \phi(\theta)$, of the full parameter θ.

I'll give a brief summary of the general ideas here, saving details for the examples in Section 7.3; see, also, Martin and Liu (2016, Ch. 6–7) and the relevant references.

Consider the case of iid data, $Y = (Y_1, \ldots, Y_n)$. There are two obvious ways to proceed. The first is to develop an association of the form (9) with $U = (U_1, \ldots, U_n)$, an auxiliary variable for every data point, introduce a random set for the n-dimensional U, and proceed as above. The second is to construct valid inferential models with output b_{y_i} for each $i = 1, \ldots, n$, and than combine—since they are all independent and carry

28

information about the same θ—combine them in some way, e.g., using Dempster's rule of combination. It turns out that neither of these two obvious approaches are most efficient; see Section 8.2. The key insight, first presented in Martin and Liu (2015a), is that if information about θ is coming from multiple sources, then there will features of the unobservable auxiliary variable $U = (U_1, \ldots, U_n)$ that actually *are observed*, hence no reason to predict those features with a random set. The strategy is to express U in a coordinate system in wherein some coordinates are observed and the others are not. This is done by identifying a one-to-one mapping $u \mapsto (\tau(u), \eta(u))$, where $\eta(u)$ corresponds to the observable part and $\tau(u)$ the unobservable part. Then $\tau(U)$ will necessarily be of lower dimension than U itself, so predicting $\tau(U)$ with a suitable random set will be easier and more efficient than predicting the full U. There are a number of strategies available to identify the (τ, η) mapping—using sufficient statistics, symmetry/invariance, a conditional representation of the statistical model, solving a differential equation, etc.— and I will illustrate these different approaches in the examples that follow.

Next, assuming that the combination steps have been carried out, so that the auxiliary variable U is currently of the same dimension as that of the parameter θ, if the goal is inference on $\phi = \phi(\theta)$, then there may be an opportunity to reduce the dimension of the auxiliary variable further. In the simplest case, imagine that some feature of U is directly connected to data and ϕ, while the remaining features of U don't directly connect to ϕ. Martin and Liu (2015b) show that one can effectively *ignore* the part of U unrelated to ϕ, and predict only that relevant feature of U. Again, that feature would be of dimension lower than that of the original U, hence an opportunity for improved efficiency. Examples below will illustrate this. Being able to isolate the interest parameter ϕ in the association is not always possible, it depends on the structure of the model, but there are strategies that may work even if the isolation is not strictly possible; see Section 7.3.3.

7.3 Examples

7.3.1 Two normal means

Consider two independent normal samples, i.e.,

$$Y_{11}, \ldots, Y_{1n_1} \overset{\text{iid}}{\sim} \mathsf{N}(\mu_1, \sigma^2) \quad \text{and} \quad Y_{21}, \ldots, Y_{2n_2} \overset{\text{iid}}{\sim} \mathsf{N}(\mu_2, \sigma^2), \tag{14}$$

where the means, μ_1 and μ_2, and the sample sizes, n_1 and n_2, are possibly different, but the variance σ^2 is assumed to be the same in both groups; the more general case is considered in Section 7.3.3. The full parameter in this case is $\theta = (\mu_1, \mu_2, \sigma)$, but the goal is only to determine if there is a difference between the two group means, a common marginal inference problem.

As a first step, according to the general results in Martin and Liu (2015a), reduce the dimension of the problem down via sufficient statistics. Then the relevant association is

$$\hat{\mu}_1 = \mu_i + \sigma n_1^{-1/2} U_1,$$
$$\hat{\mu}_2 = \mu_2 + \sigma n_2^{-1/2} U_2,$$
$$\hat{\sigma}^2 = \sigma^2 (n_1 + n_2 - 2)^{-1} U_3,$$

where $\hat{\mu}_1$ and $\hat{\mu}_2$ are the group sample means, $\hat{\sigma}^2$ is the usual pooled sample variance, U_1 and U_2 are independent standard normals, and U_3 is an independent chi-square with

$n_1 + n_2 - 2$ degrees of freedom. Interest is in $\phi = \mu_1 - \mu_2$ and manipulating the above association in an obvious way gives

$$\hat{\phi} = \phi + \hat{\sigma}\frac{n_1^{-1/2}U_1 + n_2^{-1/2}U_2}{\sqrt{(n_1 + n_2 - 2)^{-1}U_3}} \qquad \hat{\mu}_2 = \mu_2 + \sigma n_2^{-1/2}U_2,$$
$$\hat{\sigma}^2 = \sigma^2(n_1 + n_2 - 2)^{-1}U_3,$$

where $\hat{\phi} = \hat{\mu}_1 - \hat{\mu}_2$. The latter two equations do not directly involve the interest parameter ϕ, so if the goal is marginal inference on ϕ, I have no need to predict U_2 and U_3, so these can be ignored. And that particular feature of (U_1, U_2, U_3) in the first equation has a known distribution, namely, a Student-t distribution with $n_1 + n_2 - 2$ degrees of freedom. So if F denotes that distribution function, then the marginal association for ϕ can be written as

$$\hat{\phi} = \phi + \hat{\sigma}(n_1^{-1} + n_2^{-1})^{1/2}\,F^{-1}(W), \quad W \sim \mathsf{Unif}(0, 1). \tag{15}$$

If I now introduce the default random set $\mathcal{S}$ in (12), which *is* optimal in this case, then I get a corresponding random set for ϕ,

$$\Phi_x(\mathcal{S}) = \{\varphi : w_y(\varphi) \in \mathcal{S}\},$$

where $w_y(\varphi)$ is the solution of (15) for given data "y" and generic φ value, i.e.,

$$w_y(\varphi) = F\Big(\frac{\hat{\phi} - \varphi}{\hat{\sigma}\{n_1^{-1} + n_2^{-1}\}^{1/2}}\Big).$$

Then the plausibility is given by a probability with respect to the distribution $\mathsf{P}_\mathcal{S}$ of the random set $\mathcal{S}$:

$$p_y(\{\varphi\}) = \mathsf{P}_\mathcal{S}\{\Phi_y(\mathcal{S}) \ni \varphi\} = 1 - \big|2w_y(\varphi) - 1\big|. \tag{16}$$

For illustration, I consider two $(\hat{\phi}, \hat{\sigma})$ cases, one small signal-large noise and one large signal-small noise. For simplicity of comparison only, I assume that $n_1 = n_2 = n$, and I consider $n = 5$ and $n = 10$ for each of the two cases above. Figure 5 shows the plausibility contour $p_y(\{\varphi\})$ for the two cases and two sample size configurations.

First, in all four cases, the plausibility contour peaks at the estimate $\hat{\phi}$, and decays symmetrically as φ moves farther from $\hat{\phi}$. The interpretation is that the most likely value, the point estimator, is fully plausible—which is why it's used as an estimate—but values distant from the estimate should be less plausible. Second, the more informative the data are, either because $\hat{\sigma}$ is smaller or because the sample size is larger, the more narrow the plausibility curve is. This, again, agrees with the intuition that more informative data leads to more precise inference, i.e., a more narrow range of plausible values.

What is especially interesting about the plausibility function plots in Figure 5 is that the data analyst can immediately read off the answer to virtually any relevant question. First, using the horizontal reference line at 0.05, one can easily identify the 95% plausibility interval for ϕ as the set of all φ with point plausibility at least 0.05. Naturally, more informative data makes the interval narrower. And the interpretation is clear: if a value, such as $\varphi = 0$ is in the plausibility interval, then it is sufficiently plausible based on the data, otherwise it's not. Note also that the plausibility function is on the appropriate scale for reading off these intervals; compare this to a Bayesian approach where the need to evaluate tail areas via integration makes it impossible to read off the credible interval from a plot of the posterior density. Of course, there is nothing sacred about 0.05, one is free to set their own threshold for "sufficiently implausible."

(a) $\hat{\phi} = 0.5$ and $\hat{\sigma} = 1.25$.   (b) $\hat{\phi} = 1$ and $\hat{\sigma} = 0.75$.
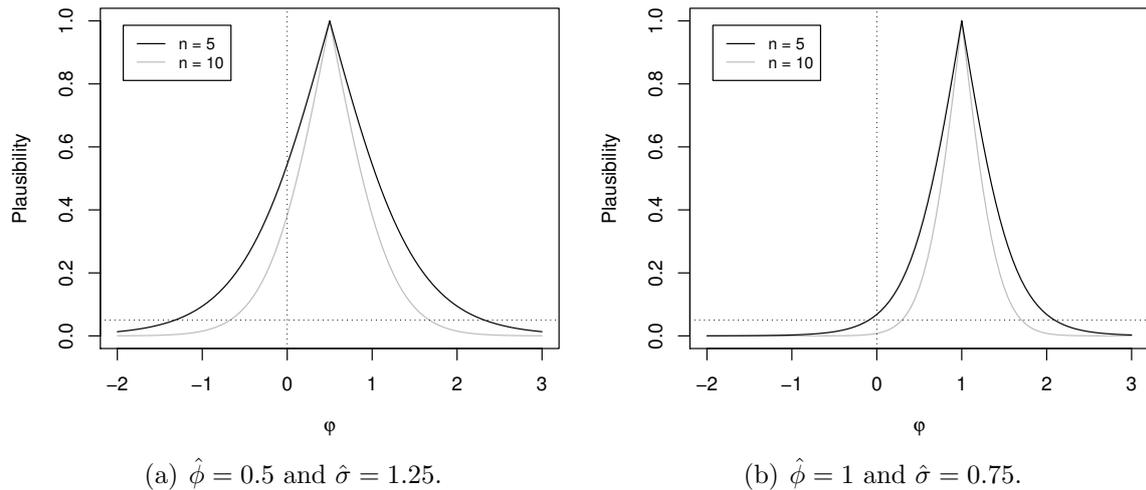
Figure 5: Plots of the plausibility contour from Section 7.3.1 for the two cases and for the two sample size configurations in each case, all on the same scale.

Second, since interest is in comparing the group means, a relevant hypothesis is $A = (-\infty, 0)$. The reference line at $\varphi = 0$ is helpful in this assessment. In Panel (a), for example, with $n = 5$ (black line), I find that $p_y(\{0\})$ is between 0.5 and 0.6; the precise value is 0.545. Therefore, since the point plausibility is increasing up to $\varphi = 0$, I get $p_y(A) = 0.545$ and $b_y(A) = 0$, so no definitive judgment can be made about the hypothesis A because belief is small but plausibility is not small. Virtually the same conclusion would be reached in the slightly more informative case with $n = 10$. On the other hand, in Panel (b), with $n = 10$ (gray line), $p_y(\{0\})$ is effectively 0, so both $b_y(A)$ and $p_y(A)$ are very small, so I can infer A^c, i.e., the hypothesis $\{\phi < 0\} = \{\mu_1 < \mu_2\}$ is sufficiently implausible based on the data.

7.3.2 Normal coefficient of variation, cont.

Reconsider the normal coefficient of variation example from Section 3.2. Following the general theory, take the minimal sufficient statistic $(\hat{\mu}, \hat{\sigma}^2)$, the sample mean and sample variance, respective, and build the association in terms of these:

$$\hat{\mu} = \mu + \sigma n^{-1/2} U_1, \qquad\qquad U_1 \sim \mathsf{N}(0, 1),$$
$$\hat{\sigma}^2 = \sigma^2 U_2, \qquad\qquad U_2 \sim (n-1)^{-1}\mathsf{ChiSq}(n-1),$$

where U_1 and U_2 are independent. The parameter of interest, $\phi = \sigma/\mu$, is a scalar but the auxiliary variable (U_1, U_2) is two-dimensional, so we would like to further reduce the dimension of the latter. An argument similar to that in Section 7.3.1 leads to a (marginal) association for the interest parameter ϕ:

$$n^{1/2}\hat{\phi}^{-1} = F_{n,\phi^{-1}}^{-1}(W), \quad W \sim \mathsf{Unif}(0, 1),$$

where $\hat{\phi} = \hat{\sigma}/\hat{\mu}$ and $F_{n,\psi}$ is the non-central Student-t distribution function, with $n-1$ degrees of freedom and non-centrality parameter $n^{1/2}\psi$. Different random sets $\mathcal{S}$ for the unobserved W may be considered (see below) but, for now, I'll consider that same "default" choice in (12). The calculation in Example 1 leads to a plausibility contour

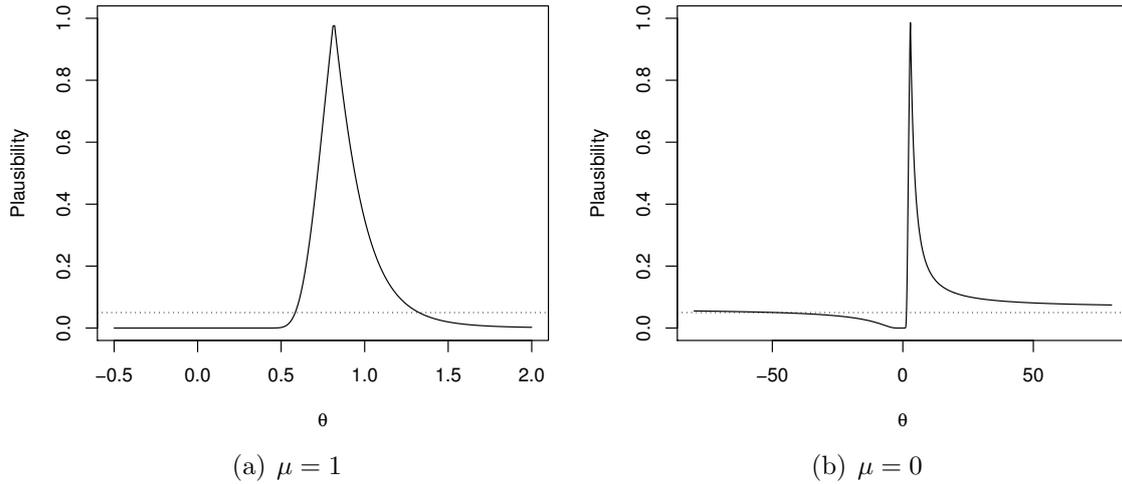$$p_y(\{\varphi\}) = 1 - |2F_{n,\varphi^{-1}}(n^{-1}\hat{\phi}^{-1}) - 1|. \tag{17}$$

31

(a) $\mu = 1$ (b) $\mu = 0$

Figure 6: Plots of the plausibility function for $\phi = \sigma/\mu$ based on samples of size $n = 30$ from $\mathsf{N}(\mu, \sigma^2)$ for $\sigma = 1$ and two values of μ.

Since the random set is nested, the derived belief function is consonant and, hence, this plausibility contour fully determines the inferential model output.

For given $\alpha \in (0, 1)$, the $100(1 - \alpha)\%$ plausibility region for ϕ is given by

$$\{\varphi : p_y(\{\varphi\}) > \alpha\} = \{\varphi : \alpha/2 < F_{n,\varphi^{-1}}(n^{1/2}\hat{\phi}^{-1}) < 1 - \alpha/2\}.$$

One can check directly that this plausibility region has coverage probability $1 - \alpha$, but it follows as a consequence of Theorem 2. The Gleser–Hwang theorem applies to these plausibility regions, so having the nominal frequentist coverage implies that these will be unbounded with positive probability; see Figure 6, which shows two plausibility contours, one that vanishes in the tails, yielding bounded plausibility intervals, and one that does not. Of course, unbounded regions are perfectly reasonable in this problem: given the inherent variability of the sample mean around μ, if μ is close to zero, then arbitrarily large values of ϕ cannot be ruled out.

Next, I revisit the simulation from Section 3.2. Recall that Y consists of $n = 10$ iid observations from $\mathsf{N}(\mu, \sigma^2)$, with $\mu = 0.1$ and $\sigma = 1$, so that $\phi = 10$ is the true value. Then I consider a false hypothesis $A = (-\infty, 9]$. Here I want to compare the distribution of the belief function $b_Y(A)$ from above, as a function of Y, with that of the posterior distribution $\Pi_Y(A)$ from before. Figure 7 shows the same distribution function of $\Pi_Y(A)$ as in Figure 2, and also shows two other distribution functions: one for the belief function $b_Y(A)$ based on the inferential model construction laid out above; the other for the belief function based on the same formulation but with a "better" random set, namely, the one-sided interval $\mathcal{S} = [\widetilde{W}, 1]$, with $\widetilde{W} \sim \mathsf{Unif}(0, 1)$, suggested by Theorem 4 in Martin and Liu (2013). Both of the latter distribution functions stay above the diagonal line, because they're both valid, while the Bayes version falls below and, hence, is not valid. That the one belief has distribution function closer to the diagonal line, without going below, is a sign of added efficiency; see Section 8.2. The takeaway message is that it's possible to construct valid (and even efficient) degrees of belief in this difficult problem, but non-additivity is essential.
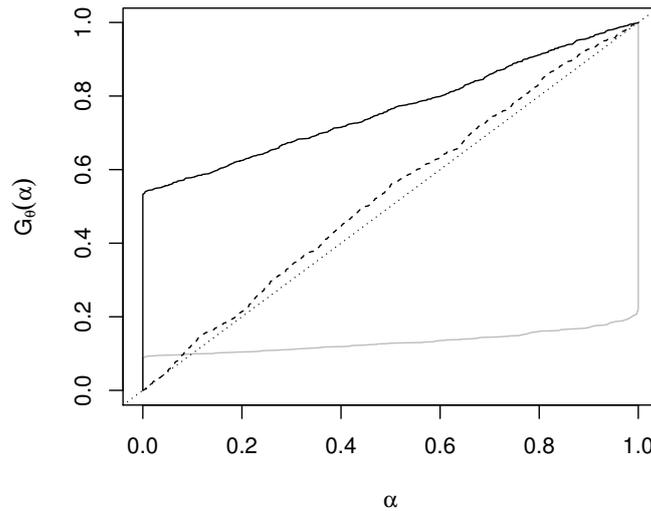
32

Figure 7: Plot of the distribution function $G_\theta(\alpha) = \mathsf{P}_{Y|\theta}\{b_Y(A) \leq \alpha\}$ of the Bayesian posterior probability (gray), the valid inferential model (black, solid) with plausibility contour in (17), and the one based on the "better" random set (dashed), where $A = (-\infty, 9]$ is a false hypothesis about the coefficient $\phi = \sigma/\mu$ when $\mu = 0.1$ and $\sigma = 1$.

7.3.3 Behrens–Fisher problem

The so-called *Behrens–Fisher problem* is a fundamental one, a generalization of the problem in Section 7.3.1 where the two groups have their own unknown variances. Despite its apparent simplicity, it turns out that the only "exact" procedures are those that depend on the order in which the data is processed, which is not fully satisfactory. Standard solutions are given by Hsu (1938) and Scheffé (1970); see, also, Welch (1938, 1947). For a review, see Kim and Cohen (1998), Ghosh and Kim (2001), and Fraser et al. (2009).

Consider the model in (14), but where the two groups have their own unknown variance, σ_1^2 and σ_2^2, respectively. Reduction via sufficiency leads to a baseline association

$$\hat{\mu}_k = \mu_k + \sigma_k \, n_k^{-1/2} \, U_{1k}, \quad \text{and} \quad \hat{\sigma}_k = \sigma_k U_{2k}, \quad k = 1, 2, \tag{18}$$

where $U_{1k} \sim \mathsf{N}(0, 1)$ and $(n_k - 1)U_{2k}^2 \sim \mathsf{ChiSq}(n_k - 1)$, independent, for $k = 1, 2$. Define $f(\sigma_1, \sigma_2) = (\sigma_1^2/n_1 + \sigma_2^2/n_2)^{1/2}$ and note that (18) can be re-expressed as

$$\hat{\phi} = \phi + f(\sigma_1, \sigma_2) V \quad \text{and} \quad \hat{\sigma}_k = \sigma_k U_{2k}, \quad k = 1, 2, \quad V \sim \mathsf{N}(0, 1). \tag{19}$$

Plug in $\hat{\sigma}_k = \sigma_k U_{2k}$ for $k = 1, 2$, to get

$$\frac{\hat{\phi} - \phi}{f(\hat{\sigma}_1, \hat{\sigma}_2)} = \frac{f(\sigma_1, \sigma_2)}{f(\sigma_1 U_{21}, \sigma_2 U_{22})} V, \quad \text{and} \quad \hat{\sigma}_k = \sigma_k U_{2k}, \quad k = 1, 2. \tag{20}$$

If the coefficient on V was free of (σ_1, σ_2), it would be straightforward to marginalize to an association in terms of ϕ only, by ignoring the nuisance parameter components. However, this simple approach is not possible in this challenging problem. As an alternative, Martin and Liu (2015b) suggest a strategy in which the nuisance parameter components can be ignored but that the distribution of the auxiliary variable is stretched accordingly.
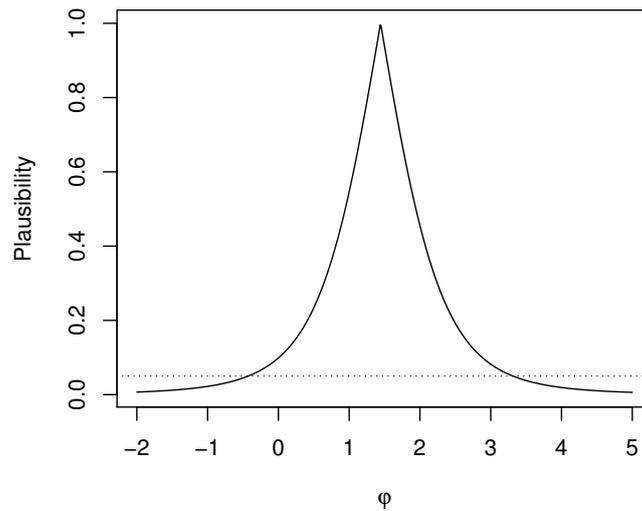
33

Figure 8: Plot of the plausibility contour for the example in Section 7.3.3. Horizontal line at $\alpha = 0.05$ determines the 95% plausibility region for ϕ.

Skipping the details, they suggest a new dimension-reduced association for ϕ,

$$\frac{\hat{\phi} - \phi}{f(\hat{\sigma}_1, \hat{\sigma}_2)} = G^{-1}(W), \quad W \sim \mathsf{Unif}(0, 1),$$

where G is the distribution function of a Student-t distribution with $\max(n_1, n_2) - 1$ degrees of freedom. Here, because of symmetry, the default random set $\mathcal{S}$ in (12) is optimal, and the plausibility contour can be evaluated exactly as in Example 1 above. Moreover, this generalized marginal inferential model is valid, which leads to an alternative proof of the conservative coverage properties of the Hsu–Scheffé confidence interval.

Data on travel times from home to work for two different routes are presented by Lehmann (1975, p. 83), summarized below:

$$(n_1, \hat{\mu}_1, \hat{\sigma}_1^2) = (5, 7.58, 2.24) \quad \text{and} \quad (n_2, \hat{\mu}_2, \hat{\sigma}_2^2) = (11, 6.14, 0.07).$$

The goal is to investigate the difference between the mean travel times for the two routes. Figure 8 shows the plausibility contour and along with the horizontal cut at $\alpha = 0.05$ that determines the 95% plausibility interval for ϕ. As expected, the contour has a peak at the estimated mean difference, $\hat{\phi} = 1.44$, but the fact that $p_y(\{0\})$ exceeds 0.05 implies that the two routes having the same mean travel times is sufficiently plausible.

7.3.4   Two Poisson rates

Consider two independent Poisson samples, $Y_1 \sim \mathsf{Pois}(n_1 \theta_1)$ and $Y_2 \sim \mathsf{Pois}(n_2 \theta_2)$, where the "sample sizes" n_1 and n_2 are known but the rates $\theta = (\theta_1, \theta_2)$ are unknown. The quantity of interest here is $\phi = \theta_1 / \theta_2$, the ratio. Inference on ϕ has been explored in a number of papers, including Krishnamoorthy and Thomson (2004) and the references therein. A similar problem arises in Jin et al. (2016), and I employed a conditioning strategy like the one below in Martin (2018, Sec. 5).

The jumping off point is the realization that there are two data points—each with a corresponding auxiliary variables—but only a scalar interest parameter. So reducing the

dimension before building the inferential model would be advantageous. One dimension-reduction strategy that can sometimes be employed is based on conditioning. Here the approach is relatively straightforward and familiar. That is, the conditional distribution of Y_2, given $Y_1 + Y_2 = k$, is binomial with parameters k and

$$g(\phi) = \frac{n_2 \theta_2}{n_1 \theta_1 + n_2 \theta_2} = \{1 + (n_1/n_2)\phi\}^{-1},$$

a one-to-one function of ϕ. So I just have to build an association in terms of the marginal distribution for $Y_1 + Y_2$ and also in terms of the aforementioned conditional. Each will have an associated auxiliary variable but I will just ignore the one in the marginal part, focusing only on the association that describes the conditional distribution, which only involves the interest parameter. And, fortunately, that analysis based on the conditional distribution is almost identical to that presented in Example 2 above. The only point of departure is the transformation between the binomial parameter $g(\phi)$ and ϕ, but this is not a problem. Here, however, I will also consider a random set other than the default considered in all the examples so far. In particular, I will use the greedy random set construction laid out in Martin et al. (2012) which, unfortunately, is too complicated to describe here; my reason for doing so is to illustrate what an *efficient* inferential model looks like in a discrete problem; see, also, Section 8.4 below.

As an illustration, suppose a builder purchases lumber from both Company 1 and Company 2. Whenever the lumber is delivered, a sample is taken and inspected and, in the most recent deliveries, a sample of $n_1 = 30$ samples was taken from Company 1's lot and $y_1 = 2$ defectives were identified, while $n_2 = 35$ samples from Company 2's delivery yielded $y_2 = 4$ defectives. Using the conditional-binomial formulation described above, with random set construction as in Martin et al. (2012), I get the plausibility contour for ϕ in Figure 9. One might guess that these data can't distinguish between the two lumber suppliers in terms of their defective rates, and the plausibility contour plot confirms this since the plausibility at $\varphi = 1$ is rather high. What's especially interesting about the plot is that it's not smooth, there are both plateaus and corners. My experience is that "good" random sets for discrete data problems will yield plausibility contours with this shape, but a clear understanding of this phenomenon is still lacking.

7.3.5   Fréchet shape

Suppose that data $Y = (Y_1, \ldots, Y_n)$ are iid from a unit Fréchet distribution with shape parameter $\theta > 0$; usually there are location and scale parameters included but I'm leaving these out here only for simplicity. The Fréchet distribution is *max-stable* and, therefore, is a canonical model in extreme value theory (e.g., de Haan and Ferreira 2006). It's also a challenging problem because it lacks the nice exponential family structure that guarantees existence of a one-dimensional sufficient statistic. So combining the different sources of information about θ in this case requires some non-standard techniques.

To construct a valid inferential model, first recall that the distribution function for the Fréchet model is given by

$$F_\theta(y) = \exp(-y^{-\theta}), \quad y > 0,$$

and so this model can be described via an association

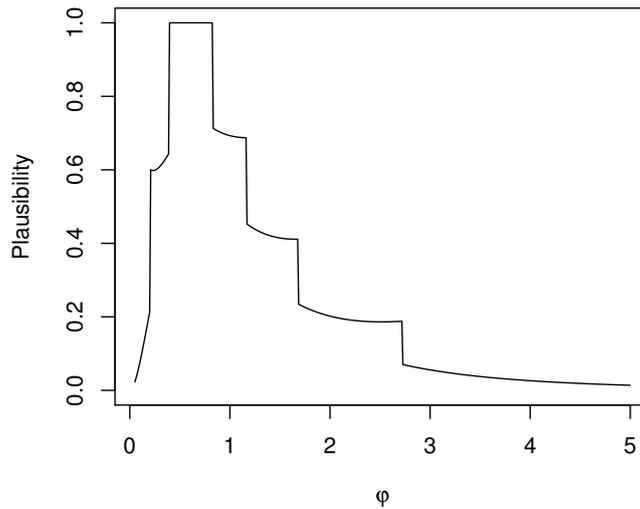$$Y_i = (-\log U_i)^{-1/\theta}, \quad i = 1, \ldots, n, \tag{21}$$

35

Figure 9: Plot of the plausibility contour for the ratio $\phi = \theta_1/\theta_2$ of Poisson means based on the data as described in Example 7.3.4.

where $U_1, \ldots, U_n$ are iid $\mathsf{Unif}(0,1)$ and the expression on the right-hand side above is just the solution $F_\theta^{-1}(u)$ of the equation $u = F_\theta(y)$. As I hinted in Section 7.2, in such cases there will be features of the auxiliary variable that are fully observed, and that identifying those features—and suitably conditioning on them—is important for reducing the dimension and increasing the efficiency of the inferential model. One might be able to guess directly what kind of features are observed (see below), but let me proceed with a formal construction which I think is informative.

The jumping off point is the observation that, if $u_{y,\theta}$ is a solution to the system (21), then those features, $\eta(u_{y,\theta})$, that don't vary with θ would correspond to observed features of U. So then the goal would be to find a function η that satisfies the following partial differential equation:

$$\frac{\partial \eta(u_{y,\theta})}{\partial \theta} = 0.$$

Presently, it is unknown what problems will admit a differential equation that has a solution, but there are examples where solving this is possible, and there are other examples (see Section **??** below) where some change-of-perspective is required in order to find a solution. For the present case with n iid samples, it makes sense to use the chain rule and re-express this differential equation as

$$\left.\frac{\partial \eta(u)}{\partial u}\right|_{u=u_{y,\theta}} \cdot \frac{\partial u_{y,\theta}}{\partial \theta} = 0, \tag{22}$$

where $\partial u_{y,\theta}/\partial \theta$ is a n-vector that is easy to compute from the known form of $u_{y,\theta}$, and $\partial \eta(u)/\partial u$ is a matrix with n columns and of rank $n-1$. That is, η is a map from the n-space of u's to some $(n-1)$-dimensional subspace.

Solving the above differential equation can proceed in various ways; techniques that I've found to be successful are guess-and-check and the *method of characteristics* (e.g., Polyanin et al. 2002) and here I'll illustrate the latter. Write $m_\theta(u)$ for $\partial u_{y,\theta}/\partial \theta$ with $y = a(\theta, u)$ plugged in, so that it's only a function of θ and u. In the present example,

$$m_\theta(u) = \theta^{-1} u \log u \log(-\log u), \quad u \in (0,1).$$

36

Then the method of characteristics suggests solving the following (nonparametric) system of ordinary differential equations:

$$\frac{du_1}{m_\theta(u_1)} = \frac{du_2}{m_\theta(u_2)} = \cdots = \frac{du_n}{m_\theta(u_n)}.$$

The solution to such a system of equations looks like

$$z_\theta(u_1) = z_\theta(u_2) + c_1 = \cdots = z_\theta(u_n) + c_{n-1}$$

for constants $c_1, \ldots, c_{n-1}$, where $dz_\theta(u)/du = m_\theta(u)^{-1}$. In the present example,

$$z_\theta(u) = \int m_\theta(u)^{-1}\, du = \theta \int \frac{1}{u \log u \log(-\log u)}\, du,$$

which, upon making a change-of-variable, $s = \log(-\log u)$, yields an anti-derivative

$$z_\theta(u) = \theta \log|\log(-\log u)|.$$

The constants $c_1, \ldots, c_{n-1}$ are intended to accommodate the rank constraint on the derivative of η, and suggests a sort of "location constraint" on the $z_\theta(u_i)$'s, so that the final solution is a function of the differences $z_\theta(u_i) - z_\theta(u_j)$. That is, take

$$\eta(u) = M\, z_\theta(u),$$

where $z_\theta(u) = (z_\theta(u_1), \ldots, z_\theta(u_n))^\top$ and M a matrix with constant vectors in its null space, so that the rows of M are like contrasts. Note that θ appearing as a multiplicative constant in $z_\theta(u)$ can be ignored here, so that $\eta(u)$ only depends on u. From this final expression, it is easy to see that η satisfies the above differential equation.

In this particular example, the differences $z_\theta(u_i) - z_\theta(u_j)$ are proportional to the logarithm of the ratio $\log(-\log u_i)/\log(-\log u_j)$, which is easily seen to be equal to the ratio $\log y_i/\log y_j$. Therefore, the feature $\eta(u)$ constructed above is observable, as intended. The reader experienced with properties of the Fréchet distribution might have been able to anticipate this form of η from the beginning, but I think this formal construction is still interesting and informative.

To keep things simple here, let me consider only the $n = 2$ case. Then I can simplify the final result above and take

$$\eta(u) = \log(-\log u_2)/\log(-\log u_1).$$

I need another mapping, τ, to complement η, and a suitable choice here is $\tau(u) = \log(-\log u_1)$. This leads to a re-expressed "dimension-reduced" association

$$T(Y) = \theta^{-1}V_1 \quad \text{and} \quad H(Y) = V_2,$$

where $T(Y) = -\log Y_1$, $H(Y) = \log Y_2/\log Y_1$, $V_1 = \tau(U)$, and $V_2 = \eta(U)$. It is straightforward to find the conditional distribution of V_1, given $V_2 = h$, where h is the observed value of $H(Y)$, and then define a random set, $\mathcal{S}$, designed to predict draws from this conditional distribution. That is, if g_h is the density for the aforementioned conditional distribution, then a very reasonable—perhaps "optimal"—random set is

$$\mathcal{S} = \{v_1 : g_h(v_1) \geq g_h(\tilde{V}_1)\}, \quad \tilde{V}_1 \sim g_h.$$

With that the plausibility contour can be evaluated, via Monte Carlo, as

$$p_y(\{\vartheta\}) \approx \frac{1}{K} \sum_{k=1}^{K} 1\{g_h(\tilde{V}_1^{(k)}) \geq g_h(\vartheta T(y))\}, \quad \tilde{V}_1^{(k)} \overset{\text{iid}}{\sim} g_h.$$

A plot of this plausibility contour, based on a pair of samples ($n = 2$) from F_θ with $\theta = 2$, is shown in Figure 10.
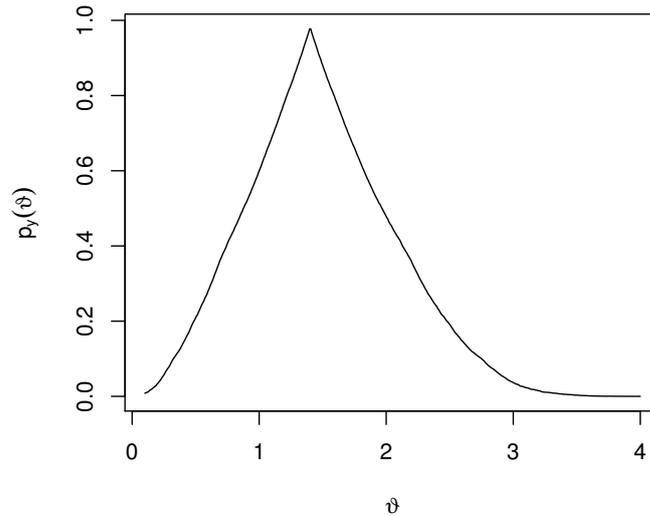
Figure 10: Plausibility contour for the Fréchet example; $n = 2$, true $\theta = 2$.

7.3.6   Non-regular cube-root normal

Consider the following *cube-root normal* problem,

$$Y_1 = \theta + U_1 \quad \text{and} \quad Y_2 = \theta^{1/3} + U_2,$$

where (U_1, U_2) consists fo two iid $\mathsf{N}(0,1)$ random variables, and the goal is inference on θ. This example was considered in Seidenfeld (1992) for comparing fiducial and Bayes updates, and more recently in Taraldsen and Lindqvist (2017). There's two observations carrying information about the same θ, so there is an opportunity to reduce the dimension of the auxiliary variable to match that of the parameter. No such reduction is provided by sufficiency in this non-regular example, but the differential equation strategy can be applied here to reduce the dimension.

Write $u_{y,\theta}$ for the solution of the above system when y and θ are fixed:

$$u_{y,\theta} = (y_1 - \theta, y_2 - \theta^{1/3})^\top.$$

As before, the goal is to find $\eta(u)$ to solve the differential equation

$$\left(\frac{\partial \eta(u)}{\partial u} \Big|_{u = u_{y,\theta}} \right)^\top \frac{\partial u_{y,\theta}}{\partial \theta} = 0.$$

Unfortunately, it may not be possible to find a solution to this differential equation that doesn't depend on θ. But there's a way to get around this obstacle based on an idea of *localization* first presented in Martin and Liu (2015a). Let $\eta_\vartheta(u)$ be a function like before, depending on some generic parameter value ϑ. Localization requires that we solve the following differential equation:

$$\frac{\partial \eta_\vartheta(u_{y,\theta})}{\partial \theta} = 0 \quad \text{when } \theta = \vartheta.$$

With this additional flexibility, the equation *can* be solved:

$$\eta_\vartheta(u) = \vartheta^{2/3} u_2 - \tfrac{1}{3} u_1.$$

Now pair this η_ϑ with a function τ such that $u \mapsto (\tau(u), \eta_\vartheta(u))$ is one-to-one for each ϑ. Here I will take $\tau(u) = u_1$. There is corresponding pair of functions (T, H_ϑ) on the Y-space, defined as

$$T(y) = y_1 \quad \text{and} \quad H_\vartheta(y) = \eta_\vartheta(u_{y,\vartheta}) = 3\vartheta^{2/3}(y_2 - \vartheta^{1/3}) - (y_1 - \vartheta).$$

Write out a "dimension-reduced" association

$$T(Y) = \theta + \tau(U) \quad \text{and} \quad H_\vartheta(Y) = \eta_\vartheta(U).$$

The first equation depends on θ while the other does not. This suggests writing $V_1 = \tau(U)$ and $V_2 = \eta_\vartheta(U)$ and using the conditional distribution of V_1, given the observed value $h_\vartheta = H_\vartheta(y_1, y_2)$ of V_2. Of course, this conditional distribution is normal and can be worked out explicitly, although it's tedious; see below. The symmetry of the normal auxiliary variable distribution suggests using the default random set in (12).

   The above construction yields a different inferential model for each value ϑ. The key insight in Martin and Liu (2015a) was to "glue" these inferential models together on the plausibility contour scale. That is, if $p_y(\cdot \mid \vartheta)$ was the plausibility contour for the ϑ-dependent inferential model, then define the glued inferential model to have contour

$$p_y(\{\vartheta\}) = p_y(\{\vartheta\} \mid \vartheta),$$

where the localization point varies with the argument to the plausibility contour. Since the individual inferential models are valid, it is easy to show that the glued version is valid as well. This gluing was used in other non-regular examples, including those in Martin and Lin (2016) and in Cheng et al. (2014). In this example, with the default random set, the glued plausibility contour is

$$p_y(\{\vartheta\}) = 1 - \left| 2\Phi\Big(\frac{y_1 - \vartheta + \frac{1}{3\vartheta^{2/3}}(y_2 - \vartheta^{1/3})}{(1 + \frac{1}{9\vartheta^{4/3}})^{1/2}}\Big) - 1 \right|.$$

For illustration, I simulate data from the above model with $\theta = 7$. Figure 11 shows a plot of this glued plausibility contour, along with those obtained by treating y_1 and y_2 separately. Since y_1 is more informative about θ than y_2, it makes sense that the combined plausibility contour gives that observation more weight.

8   More remarks

8.1   A new take on the fiducial argument

In Section 5.1 I described Fisher's version of the fiducial argument, circa 1930. That argument is delightful, but requires too much structure. Dempster (1963, 1964) describes it in a different way, which I'll summarize here and then relate back to what's going on the valid inferential model construction.

   Consider the baseline association (9) from which data can be generated, i.e., sample $U \sim \mathsf{P}_U$, plug into the mapping, for a given θ, and return $Y = a(\theta, U)$. Clearly, U and Y are dependent; in fact, they're very tightly linked, via a deterministic mapping, so there's no joint distribution as usual. If one had a "conditional distribution of U, given $Y = y$," then one could just sample, say $\tilde{U}$, from that distribution, plug it into $y = a(\theta, \tilde{U})$, and solve for θ to get a draw from its "posterior distribution." The problem, however, is
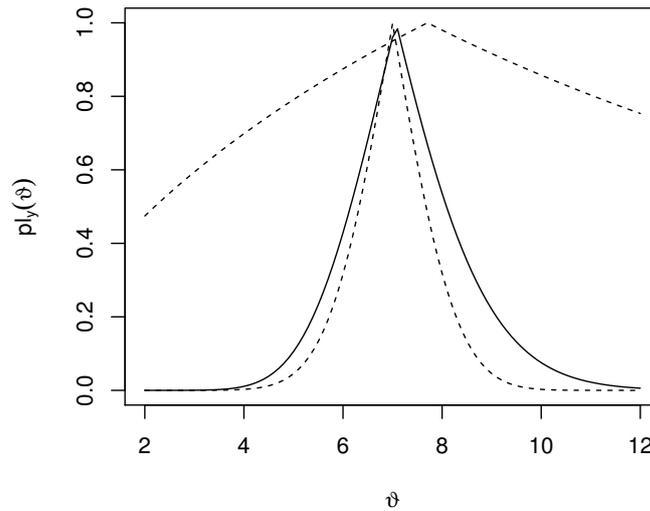
Figure 11: Plausibility contour for cube-root normal problem, true $\theta = 7$. Dashed lines are based on the individual data points, while the solid line is after combination using the differential equation strategy.

that there's no such conditional distribution, at least not one that's useful. That is, if the association mapping a is sufficiently nice, then observing $Y = y$ immediately fixes U at $u_{y,\theta}$, the solution to the equation $y = a(\theta, u)$, where θ is the true value. In other words, $\mathsf{P}_{U|Y=y} = \delta_{u_{y,\theta}}$, where δ_x denotes a point mass distribution at x. Of course, this "conditional distribution" is useless because it depends on the true-but-unknown θ.

If not from probability calculus, where then does the fiducial distribution come from? The word "fiducial" means *based on trust*, and Dempster made Fisher's leap of faith clear. Indeed, Fisher just defines—or assumes—a particular conditional distribution, namely,

$$\mathsf{P}_{U|Y=y} \overset{\text{set}}{=} \mathsf{P}_U.$$

As Dempster described it, Fisher's suggestion was to "continue to regard" U as if its distribution where P_U even after observing $Y = y$, effectively ignoring the strong dependence between them. Fisher's "continue to regard" idea is clever, and has been used for many years by those working in the fiducial domain.

The point here so far is that, no surprise, it's not possible to construct a genuine posterior/conditional distribution out of thin air. And if the fiducial distribution isn't built up according to the rules of probability, then it's again no surprise that it can break down when one puts the pressure of the probability calculus on it.

At a high level, what causes problems in Fisher's argument is the fact that Y and U are too tightly tied together for ordinary probability calculus to handle. That is, if U and Y are deterministically linked, as in (9), in a θ-dependent way, then conditioning won't work. The question, then, is how to break that tight link? Dempster's efforts, in a series of classic papers (Dempster 1966, 1967, 1968a,b, 1969), to develop what he now calls a *state space model* that describes all the unknowns and the various manipulations (e.g., Dempster's rule of combination) are carried out is, I think, in this vein.

The approach described in Section 7 has similar motivation and I want to focus the rest of this subsection on that. As described above, the difficulties with fiducial arise because there's too tight of a link between Y and U. On the other hand, our original

40

inspiration for the above inferential model construction was this tight connection. That is, if identification of $u_{y,\theta}$ were possible, then the true θ could be similarly identified. So the tight link between Y and U makes it possible to convert the problem of inference about θ to one of predicting or guess $u_{y,\theta}$. This new task might not seem any easier than the previous, but the difference is that there's information available, in the generator P_U, about roughly where $u_{y,\theta}$ is located. That information is used to generate random sets—or "nets"—to be cast in order to catch the fixed $u_{y,\theta}$. There's no longer any concern about the link between Y and U because these random sets live on an entirely different probability space, independent of data and original auxiliary variable. Since the target $u_{y,\theta}$ is a draw from P_U, it makes sense to design these random sets that they can hit targets drawn from P_U with some amount of frequency. If that can be arranged, then the set of solutions ϑ to $y = a(\vartheta, u)$ corresponding to u in that set will inherit exactly the same frequency of containing the true θ, and that's basically the proof of Theorem 3. Of course, I would also have the option to choose the random set to be a singleton filled with a draw from P_U, like Fisher's and Dempster's solutions, but this typically will not satisfy the desired validity property. See Liu and Martin (2015).

Therefore, the construction presented in Section 7 above is very much in line with Fisher's fiducial argument, most importantly, because the goal is to construct degrees of belief about the parameter where none had existed before, at least not formally. The key difference, however, is that the inferential model construction doesn't simply settle with whatever comes out of the "continue to regard" step, it proactively seeks out a random set formulation with the desired validity property.

8.2   Efficiency and optimality

In the above examples, I have primarily made use of the default random set $\mathcal{S}$ defined in (12). In some examples, like the gamma model in Example 1, this choice was for simplicity, but in others I claimed that the default $\mathcal{S}$ was "optimal" in some sense. Here I want to say a few words about *efficiency* in the inferential model context, and the corresponding notion of *optimality*, which means "most efficient."

In classical statistics, efficiency can be expressed in terms of small (asymptotic) variance, e.g., between two confidence intervals that both attain the nominal coverage probability, the narrower one is more efficient. Recall that the size of the plausibility region (5) is controlled by the spread of plausibility contour which, in turn, is controlled by the random set $\mathcal{S}$. So, in the inferential model context, efficiency will be characterized or determined by the random set $\mathcal{S}$ or, rather, its distribution $\mathsf{P}_{\mathcal{S}}$.

For the moment, I will write $b_y(\cdot; \mathcal{S})$ to indicate that the random set $\mathcal{S}$ was used in the construction of the inferential model. Then I'll say that $\mathcal{S}$ is at least as efficient as $\mathcal{S}'$ at hypothesis A, subject to both being valid, if $b_y(A; \mathcal{S}) \geq b_y(A; \mathcal{S}')$ for all y.[15] This implies that $b_Y(A; \mathcal{S})$ is stochastically no smaller than $b_Y(A; \mathcal{S}')$ or, in other words, the spread of $p_Y(\vartheta; \mathcal{S})$ would tend to be narrower than that of $p_Y(\vartheta; \mathcal{S}')$.

A first basic result, presented in Martin and Liu (2013), is a complete-class theorem which says that, for any random set $\mathcal{S}'$, there exists a *nested* random set $\mathcal{S}$ that is at least as efficient. This explains why the inferential output is always a possibility function and not a general belief function. A nested random set has a fixed center, or core, so only

---

[15]Asking for the inequality to hold uniformly in y is indeed a strong type of efficiency, perhaps too strong. Replacing this inequality with a distributional inequality is possible, and probably better. The one result quoted here, however, holds under this strong condition.

its size and shape can vary. But once the overall shape is fixed, the size is determined by the need to meet the validity condition. To see this, consider, like in the discussion around (11), a function $h : \mathbb{U} \to \mathbb{R}$ and a random set defined by

$$\mathcal{S} = \{u \in \mathbb{U} : h(u) \leq s(\tilde{U})\}, \quad \tilde{U} \sim \mathsf{P}_U,$$

where $s : \mathbb{U} \to \mathbb{R}$ is another function that controls the size of $\mathcal{S}$. If s isn't at least as large as h everywhere, then validity might fail, and if $s > h$ uniformly, then $\mathcal{S}$ is too large, so the best choice is to take $s = h$. It is in this sense that the size of the random set is determined by its shape. So the choice of an efficient $\mathcal{S}$ boils down to choosing the appropriate shape. In certain one-dimensional situations, the optimal shape is known (Martin and Liu 2013). For example, in symmetric location problems, the optimal random set is itself symmetric, hence (12) is optimal when U is $\mathsf{Unif}(0, 1)$.

Efforts to understand what makes a "good" random set are ongoing, but my current intuition is that the support of $\mathcal{S}$, i.e., the collection of possible realizations, should match the contours of the distribution P_U, like in Section 7.3.5. What makes formalization of this intuition non-trivial is that the association mapping $a : \Theta \times \mathbb{U} \to \mathbb{Y}$ plays a role, as does the particular hypothesis $A \subset \Theta$ that one is interested in.

8.3   Practicalities

There are certainly some differences—philosophical or otherwise—between what has been presented above and the more traditional statistical analyses, but the practical differences are not as wide as one might think. Indeed, in those classical examples, such as linear regression, analysis of variance, etc., the recommended inferential model will have plausibility contour that yields hypothesis tests and confidence intervals that agree with those presented in the standard textbooks. So, very little in terms of day-to-day operations of statistics requires a fundamental change.

From an educational point of view, at least for those courses intended for non-statistics majors in social or biological sciences, instructors could decide if introducing the plausibility contour used to produce the standard hypothesis tests and confidence intervals is appropriate for their students. But I firmly believe that there is an advantage to using language like *plausibility* when communicating statistical ideas to these students compared to the more common *probability, confidence*, etc. The reason is that these students know just enough probability to be dangerous. For example, if a student confuses p-value for a probability, as many do, then he is likely to incorrectly interpret a small p-value as direct support for the truthfulness of the alternative hypothesis when, in reality, small p-value is only indirect support for the alternative, i.e., plausibility of the alternative is large but its belief could be small. That student may not be able to understand the sub-additivity properties of plausibility functions, but if he knows that a p-value represents plausibility instead of probability, and that the former is a different object that can't be manipulated by the same rules as the latter, then he's likely not to make the aforementioned mistake. And in light of the results summarized in Section 8.4, "confidence" is (mathematically) in direct correspondence to plausibility, so there is no need to use the former word if the latter is both satisfactory and sufficiently general. For a different set of students, with more of a mathematical background, ideas about how to carry out the entire intermediate-level statistics theory course based on plausibility contours are presented in Martin (2017b).

The majority of real-life data analyses consist of applying those standard methods taught in the course described above, so little or nothing would change for practicing statisticians. And it would be straightforward to modify the existing statistical software programs to return a plot of the plausibility contour if users request it.

That (certain features of) the recommended inferential model agree with traditional analyses in some problems, of course, does not imply that the solutions would agree in all problems. Indeed, there are already a number of examples where the inferential model constructed as above produce different and usually better solutions than other methods. More on the potential of valid inferential models is presented in Section 9.

8.4   The scope of valid beliefs

My main focus here has been on the construction of a valid inferential model, from which hypothesis tests and confidence regions with the desired frequentist error rate control can be immediately read off, as in Theorem 2. But, as is well known, one can get these decision procedures more directly, without full specification of degrees of belief. In fact, this is a major advantage of a "frequentist approach," the ability to break a problem down and jump directly to calibrated inferences on the quantity of interest.[16] There's been lots of work over the years developing good—and in some cases optimal—tests and confidence regions so, for those researchers who prefer to focus on procedures with good properties, what does a valid inferential model have to offer?

There are results available now which serve as a sort of converse to Theorem 2 above. Roughly speaking, if there exists a test or confidence region that controls the frequentist error rate at any specified level α, then there exists valid inferential model—potentially one that's *glued* in the sense of Section 7.3.6—whose plausibility function yields a test or confidence region at least as good (in terms of efficiency) as the given one. The first such result was presented in Martin and Liu (2014) in the context of p-values and hypothesis testing. This justifies my strategy, described in Section 8.3, to use the language of *plausibility* when interpreting p-values. That is, it's not just a superficial similarity between the two, there is a formal mathematical correspondence.

My more recent results (Martin 2017a) on the connection between valid inferential models and confidence regions go further. I show that, given a confidence region for a feature $\phi = \phi(\theta)$ of the full parameter, there exists a valid inferential model for θ whose corresponding marginal plausibility region for ϕ is at least as efficient as the given one. Moreover, there's an algorithm for carrying out this construction. As a quick illustration, consider again the binomial problem in Example 2. A classical procedure in the literature is the Clopper–Pearson confidence interval (e.g., Brown et al. 2001), and it is possible to construct a plausibility contour based on these intervals, which looks a lot like the plausibility function in Figure 4(a). Using the framework in Martin (2017a), I can convert the Clopper–Pearson intervals into a valid inferential model, and its plausibility contour is narrower, see Figure 12, hence the derived plausibility intervals are more efficient.

From Section 1, calibration is essential for reliable statistical inference. This reliability can be achieved by working with a valid inferential model or, to a certain extent, by developing tests and confidence intervals with frequentist error rate control for each individual task. The take-away message here is that there's virtually no loss of flexibility

---

[16]I've recently been working on an approach by which one can construct marginal posteriors about parameters of interest directly, without a full posterior first. We've been calling these *Gibbs posteriors* in Syring and Martin (2016, 2017a,b).
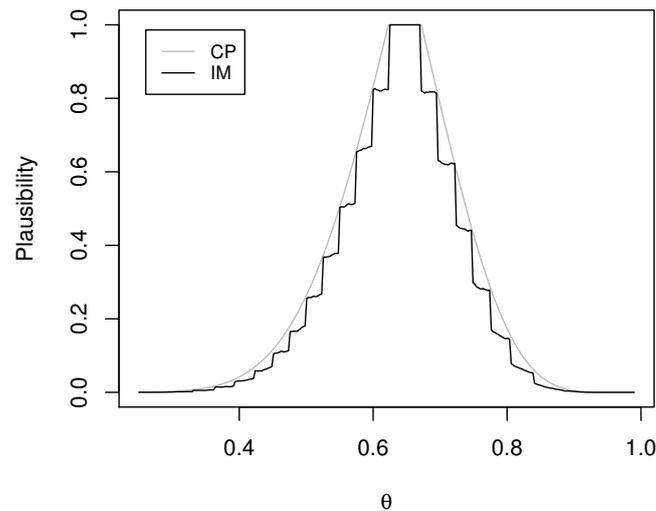
Figure 12: Plausibility contours corresponding to the Clopper–Pearson confidence interval for the binomial problem discussed in Section 8.4, along with that derived from the valid inferential model construction in Martin (2017a); $n = 20$ and $y = 13$.

or efficiency in taking the former perspective since it is always possible to construct an inferential model whose tests and confidence regions are at least as efficient.

9   Conclusion

In this paper I have presented a summary of my current views on statistical inference and its foundations. The key foundational point here is that, if one agrees with Reid and Cox that "calibrated inferences seem essential," then additive degrees of belief are not appropriate for describing uncertainty. That is, additive beliefs are afflicted with false confidence and, therefore, at least for some hypotheses, Theorem 1 says that they are prone to lead to erroneous conclusions. Two examples in Section 3 showed that the hypotheses afflicted need not be trivial ones either, so there is a genuine risk. For those data analysts who are reluctant to move away from their familiar additive beliefs, it's important to be able to identify which hypotheses are at greatest risk of false confidence so they can mitigate their damages. Towards this, those "impossible" problems highlighted in Section 3.2 should hold some insights.

My focus here has been on the situation where no prior information is available, which might be unrealistic. Is *nothing* really known about θ? What's typically done is either the available information is ignored, because it falls short of being a full prior distribution, or the inevitable information gaps are filled in by something artificial, e.g., default priors. An ideal situation is that one can incorporate *exactly* the available information, no omissions and no additions. Fortunately, there are ways in which this information can be incorporated into the inferential model framework. Recently, Qiu et al. (2018) developed a valid inferential model for the case where prior information is available about nuisance parameters, in the form of a genuine Bayesian prior distribution, but not about the parameter of interest. A student and I are currently working on a different angle, one where potentially incomplete prior information is available and to be used but, through

the combination mechanism, will only lead to an improved inferential model, i.e., validity will be maintained and no efficiency will be lost, regardless of whether the prior is "right" or "wrong," but efficiency will be gained when the prior is "right." The results so far are promising but still rather limited.

Another interesting direction I'm currently pursuing, one that has a lot in common with the partial prior problem described above, is that of *model uncertainty*. For example, suppose I am entertaining a collection of possible models, indexed by $M \in \mathcal{M}$, each having a model parameter θ_M. A Bayesian framework is equipped to handle such a problem, at least in principle, but difficulties quickly arise. For example, one might have some vague prior information about the model itself, e.g., a quantification of Occam's razor which puts more weight on simpler models, but rarely does one have genuine prior information about the individual θ_M's, and the usual rules about non-informative priors don't apply when multiple models are under consideration. An idea along these lines is to carry out the marginalization steps as described in Section 7.2 above to isolate M and, if available, introduce a sort of "prior" for M along the lines described in the previous paragraph. Work is in progress and I expect to report on this soon.

That the inferential model constructed in Section 7 is able to achieve the validity property is exciting, but since there's no free lunch, there must also be some limitations. At present, there are still many questions about this framework that I don't know how to answer, but I think this is more of an opportunity than a limitation. As far as limitations are concerned, I'll mention two here. First, since validity and efficiency are at odds with one another, and since validity is inflexible, some loss of efficiency is possible. That is, if I construct a generic inferential model about θ, then it's possible that the inference based on $b_y(A)$ or $p_y(A)$ for some specific A is not as efficient as that based on some other approach or method that targets A specifically. If, however, the inferential model is tailored to the specific hypothesis (or hypotheses) of interest, then my experience is that efficiency can be maintained. That leads to the second limitation, namely, difficulty. Presently, it is non-trivial in some cases to carry out the necessary dimension reduction steps to achieve both validity and efficiency. I will continue to think about ways to automate the process in certain ways, it's just hard to imagine what that "black box" would look like. The Bayesian approach faced similar challenges, and its present popularity is entirely due to the advent of Monte Carlo methodology that virtually automated the computations that had previously been impossible. What will the next breakthrough be? Is a similar automation of inferential model construction and belief function computation possible?

Now is an exciting time for statisticians and data scientists: lots of job opportunities in both industry and academia, interesting ("big-data") research problems with important real-life consequences, and a new generation of STEM students eager to be trained. But there are also dangers. As problems become bigger and more complex, we may reach a point where our intuition is no longer reliable. For example, dependent data is commonplace in certain applications, and dealing with that dependence is by now familiar. But modern network problems bring an entirely different meaning to "dependent data," i.e., *the data is the dependence*. Can we really expect our intuitions, developed largely based on asymptotic approximations in relatively simple cases, to safely guide us through these new complex problems? After all, false confidence is lurking in this problem too, and it's surely even more difficult to spot amidst all the complexity.

I'd like to end, just like I began, by quoting from Barnard (1985):

> *It seems to be useful for statisticians generally to engage in retrospection at this time, because there seems now to exist an opportunity for a convergence*

of view on the central core of our subject.

Before having put together the results presented in Balch et al. (2017) and the valid inferential model construction in Martin and Liu (2016), I would've said that this "convergence of view" ship had sailed. The difference is that now we understand what can go wrong in practice with additivity, not just philosophies and generalities, along with a normative approach for the construction of valid inferential models.

Acknowledgments

References

Aldrich, J. (2000). Fisher's "Inverse Probability" of 1930. *Int. Statist. Rev.*, 68(2):155–1172.

Aldrich, J. (2008). R. A. Fisher on Bayes and Bayes' theorem. *Bayesian Anal.*, 3(1):161–170.

Balch, M. S. (2012). Mathematical foundations for a theory of confidence structures. *Internat. J. Approx. Reason.*, 53(7):1003–1019.

Balch, M. S., Martin, R., and Ferson, S. (2017). Satellite conjunction analysis and the false confidence theorem. Unpublished manuscript, `arXiv:1706.08565`.

Barnard, G. A. (1985). A coherent view of statistical inference. Technical report, Department of Statistics and Actuarial Science, University of Waterloo.

Barnard, G. A. (1995). Pivotal models and the fiducial argument. *Int. Statist. Rev.*, 63(3):309–323.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour*.

Berger, J. O., Liseo, B., and Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.*, 14(1):1–28.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: irreconcilability of P values and evidence. *J. Amer. Statist. Assoc.*, 82(397):112–139.

Berger, J. O. and Wolpert, R. L. (1984). *The Likelihood Principle*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 6. Institute of Mathematical Statistics, Hayward, CA.

Bertanha, M. and Moreira, M. J. (2018). Impossible inference in econometrics: Theory and applications. Unpublished manuscript, `arXiv:1612.02024`.

Birnbaum, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.*, 57:269–326.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc. Ser. A*, 143(4):383–430. With discussion.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statist. Sci.*, 16(2):101–133. With comments and a rejoinder by the authors.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., and Wu, H. (2018). Evaluating the replicability of social science experiments in *nature* and *science* between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644.

Carmichael, I. and Williams, J. P. (2018). An exposition of the false confidence theorem. *Stat*, 7(1):e201.

Carnap, R. (1962). *Logical Foundations of Probability*. Second edition. The University of Chicago Press, Chicago, Ill.

Cheng, Q., Gao, X., and Martin, R. (2014). Exact prior-free probabilistic inference on the heritability coefficient in a linear mixed model. *Electron. J. Stat.*, 8(2):3062–3076.

Chiang, A. K. L. (2001). A simple general method for constructing confidence intervals for functions of variance components. *Technometrics*, 43(3):356–367.

Choquet, G. (1953–1954). Theory of capacities. *Ann. Inst. Fourier, Grenoble*, 5:131–295 (1955).

Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Crane, H. (2018a). The fundamental principle of probability: Resolving the replication crisis with skin in the game. *Researchers.One*, under review, `www.researchers.one/article/2018-08-16`.

Crane, H. (2018b). The impact of p-hacking on "Redefine Statistical Significance". *Basic Appl. Soc. Psych.*, to appear; `https://psyarxiv.com/bp2z4`.

Crane, H. (2018c). Polls, pundits, or prediction markets: An assessment of election forecasting. *Researchers.One*, under review, `www.researchers.one/article/2018-11-6`.

Crane, H. (2018d). *Probabilistic Foundations of Statistical Network Analysis*, volume 157 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL.

Crane, H. and Martin, R. (2018a). Academia's case of Stokholm syndrome. *Quillette*. `https://quillette.com/2018/11/29/academias-case-of-stockholm-syndrome/`.

Crane, H. and Martin, R. (2018b). In peer review we (don't) trust: How peer review's filtering poses a systemic risk to science. *Researchers.One*, under review, `www.researchers.one/article/2018-09-17`.

Crane, H. and Martin, R. (2018c). Is statistics meeting the needs of science? *Researchers.One*, under review, `www.researchers.one/article/2018-08-14`.

Crane, H. and Martin, R. (2018d). The RESEARCHERS.ONE mission. *Researchers.One*, `www.researchers.one/article/2018-07-1`.

Crane, H. and Martin, R. (2018e). Rethinking probabilistic prediction: lessons learned from the 2016 U.S. presidential election. *Researchers.One*, under review, `www.researchers.one/article/2018-08-12`.

de Finetti, B. (1937). La prévision : ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré*, 7(1):1–68.

de Finetti, B. (1972). *Probability, Induction and Statistics*. John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.

de Finetti, B. (1990). *Theory of Probability. Vol. 1*. Wiley Classics Library. John Wiley & Sons, Ltd., Chichester. A critical introductory treatment, Translated from the Italian and with a preface by Antonio Machì and Adrian Smith, With a foreword by D. V. Lindley, Reprint of the 1974 translation.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York. An Introduction.

Dempster, A. P. (1963). Further examples of inconsistencies in the fiducial argument. *Ann. Math. Statist.*, 34:884–891.

Dempster, A. P. (1964). On the difficulities inherent in Fisher's fiducial argument. *J. Amer. Statist. Assoc.*, 59:56–66.

Dempster, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Statist.*, 37:355–374.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.*, 38:325–339.

Dempster, A. P. (1968a). A generalization of Bayesian inference. (With discussion). *J. Roy. Statist. Soc. Ser. B*, 30:205–247.

Dempster, A. P. (1968b). Upper and lower probabilities generated by a random closed interval. *Ann. Math. Statist.*, 39:957–966.

Dempster, A. P. (1969). Upper and lower probability inferences for families of hypotheses with monotone density ratios. *Ann. Math. Statist.*, 40:953–969.

Dempster, A. P. (2008). The Dempster–Shafer calculus for statisticians. *Internat. J. Approx. Reason.*, 48(2):365–377.

Dempster, A. P. (2014). Statistical inference from a Dempster–Shafer perspective. In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L., editors, *Past, Present, and Future of Statistical Science*, chapter 24. Chapman & Hall/CRC Press.

Denœux, T. and Li, S. (2018). Frequency-calibrated belief functions: review and new insights. *Internat. J. Approx. Reason.*, 92:232–254.

Dubois, D. and Prade, H. (1988). *Possibility Theory*. Plenum Press, New York, french edition. An approach to computerized processing of uncertainty, With the collaboration of Henri Farreny, Roger Martin-Clouaire and Claudette Testemale, Translated by E. F. Harding, With a foreword by Lotfi A. Zadeh.

Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, 65(6):1365–1387.

Edwards, A. W. F. (1997). What did Fisher mean by "inverse probability" in 1912–1922? *Statist. Sci.*, 12(3):177–184.

Efron, B. (1998). R. A. Fisher in the 21st century. *Statist. Sci.*, 13(2):95–122.

Efron, B. (2013a). Bayes' theorem in the 21st century. *Science*, 340(6137):1177–1178.

Efron, B. (2013b). Discussion: "Confidence distribution, the frequentist distribution estimator of a parameter: a review" [mr3047496]. *Int. Stat. Rev.*, 81(1):41–42.

Ermini Leaf, D. and Liu, C. (2012). Inference about constrained parameters using the elastic belief method. *Internat. J. Approx. Reason.*, 53(5):709–727.

Evans, M. (2013). What does the proof of Birnbaum's theorem prove? *Electron. J. Stat.*, 7:2645–2655.

Evans, M. (2015). *Measuring Statistical Evidence Using Relative Belief*. Monographs in Statistics and Applied Probability Series. Chapman & Hall/CRC Press.

Feynman, R. P. (1956). The Relation of Science and Religion. Transcript of a talk given at the Caltech YMCA Lunch Forum on May 2, 1956; available at `http://calteches.library.caltech.edu/49/2/Religion.htm`.

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference.* Hafner Press, New York, 3rd edition.

Fraser, D. A. S. (1968). *The Structure of Inference.* John Wiley & Sons Inc., New York.

Fraser, D. A. S. (2011a). Is Bayes posterior just quick and dirty confidence? *Statist. Sci.*, 26(3):299–316.

Fraser, D. A. S. (2011b). Rejoinder: "Is Bayes posterior just quick and dirty confidence?". *Statist. Sci.*, 26(3):329–331.

Fraser, D. A. S. (2014). Why does statistics have two theories? In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L., editors, *Past, Present, and Future of Statistical Science*, chapter 22. Chapman & Hall/CRC Press.

Fraser, D. A. S., Bédard, M., Wong, A., Lin, W., and Fraser, A. M. (2016). Bayes, reproducibility and the quest for truth. *Statist. Sci.*, 31(4):578–590.

Fraser, D. A. S., Reid, N., Marras, E., and Yi, G. Y. (2010). Default priors for Bayesian and frequentist inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(5):631–654.

Fraser, D. A. S., Wong, A., and Sun, Y. (2009). Three enigmatic examples and inference from likelihood. *Canad. J. Statist.*, 37(2):161–181.

Fricker Jr., R. D., Burke, K., Han, X., and Woodall, W. H. (2018). Assessing the effect of *Basic and Applied Social Psychology*'s ban on p-values and hypothesis testing. Unpublished manuscipt.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis.* Chapman & Hall/CRC, Boca Raton, FL, second edition.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica*, 6(4):733–807. With comments and a rejoinder by the authors.

Ghosh, M. and Kim, Y.-H. (2001). The Behrens-Fisher problem revisited: a Bayes-frequentist synthesis. *Canad. J. Statist.*, 29(1):5–17.

Gleser, L. J. and Hwang, J. T. (1987). The nonexistence of $100(1 − \alpha)\%$ confidence sets of finite expected diameter in errors-in-variables and related models. *Ann. Statist.*, 15(4):1351–1362.

Gong, R. and Meng, X.-L. (2017). Judicious judgment meets unsettling updating: Dilation, sure loss, and Simpson's paradox. Unpublished manuscript, `arXiv:1712.08946`.

Grünwald, P. (2018). Safe probability. *J. Statist. Plann. Inference*, 195:47–63.

Hannig, J. (2009). On generalized fiducial inference. *Statist. Sinica*, 19(2):491–544.

Hannig, J. (2013). Generalized fiducial inference via discretization. *Statist. Sinica*, 23(2):489–514.

Hannig, J., Iyer, H., Lai, R. C. S., and Lee, T. C. M. (2016). Generalized fiducial inference: a review and new results. *J. Amer. Statist. Assoc.*, 111(515):1346–1361.

Hannig, J., Iyer, H., and Patterson, P. (2006). Fiducial generalized confidence intervals. *J. Amer. Statist. Assoc.*, 101(473):254–269.

Hannig, J. and Lee, T. C. M. (2009). Generalized fiducial inference for wavelet regression. *Biometrika*, 96(4):847–860.

Hedayat, A. S. and Sinha, B. K. (1991). *Design and Inference in Finite Population Sampling.* John Wiley & Sons Inc., New York.

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.*, 80:470–501.

Hinkelmann, K. and Kempthorne, O. (2008). *Design and Analysis of Experiments. Vol. 1.* Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition. Introduction to experimental design.

Hsu, P. L. (1938). Contributions to the theory of "Student's" t-test as applied to the problem of two samples. In *Statistical Research Memoirs*, pages 1–24. University College, London.

Huber, P. J. (1973). The use of Choquet capacities in statistics. *Bull. Inst. Internat. Statist.*, 45(book 4):181–191. With discussion by P. Gänssler, W. S. Hall, L. Schmetterer, Charles Stein, and G. N. Wilkinson and with reply by the author.

Huber, P. J. (1981). *Robust Statistics.* John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Huber, P. J. and Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. *Ann. Statist.*, 1:251–263.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(e124).

Ionides, E. L., Giessing, A., Ritov, Y., and Page, S. E. (2017). Response to the ASA's Statement on p-Values: Context, Process, and Purpose. *Amer. Statist.*, 71(1):88–89.

Jeffreys, H. (1961). *Theory of Probability.* Third edition. Clarendon Press, Oxford.

Jin, H., Li, S., and Jin, Y. (2016). The IM-based method for testing the non-inferiority of odds ratio in matched-pairs design. *Statist. Probab. Lett.*, 109:145–151.

Kadane, J. B. (2011). *Principles of Uncertainty.* Texts in Statistical Science Series. CRC Press, Boca Raton, FL. `http://uncertainty.stat.cmu.edu`.

Kim, S.-H. and Cohen, A. S. (1998). On the Behrens-Fisher problem: A review. *Journal of Educational and Behavioral Statistics*, 23(4):356–377.

Kohlas, J. and Monney, P.-A. (1995). *A Mathematical Theory of Hints*, volume 425 of *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Berlin. An approach to the Dempster-Shafer theory of evidence.

Krishnamoorthy, K. and Thomson, J. (2004). A more powerful test for comparing two Poisson means. *J. Statist. Plann. Inference*, 119(1):23–35.

Lai, R. C. S., Hannig, J., and Lee, T. C. M. (2015). Generalized fiducial inference for ultrahigh dimensional regression. *J. Amer. Statist. Assoc.*, 110:760–772.

Lakatos, I. (1978). *Philosophical Papers. Vol. 1.* Cambridge University Press, Cambridge-New York. The methodology of scientific research programmes, Edited and with an introduction by John Worrall and Gregory Currie.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., Cross, E. S., Daniels, S., Danielsson, H., DeBruine, L., Dunleavy, D. J., Earp, B. D., Feist, M. I., Ferrell, J. D., Field, J. G., Fox, N. W., Friesen, A., Gomes, C., Gonzalez-Marquez, M., Grange, J. A., Grieve, A. P., Guggenberger, R., Grist, J., van Harmelen, A., Hasselman, F., Hochard, K. D., Hoffarth, M. R., Holmes, N. P., Ingre, M., Isager, P. M., Isotalus, H. K., Johansson, C., Juszczyk, K., Kenny, D. A., Khalil, A. A., Konat, B., Lao, J., Larson, E. G., Lodder, G. M. A., Lukavsky, J., Madan, C. R., Manheim, R. J., Martin, S. R., Martin, A. E., Mayo, D. G., McCarthy, R. J., McConway, K., McFarland, C., Nio, A. Q. X., Nilsonne, G., de Oliveira, C., de Xivry, J. O., Parsons, S., Pfuh, G., Quinn, K. A., Sakon, J. J., Saribay, S. A., Schneider, I. K., Selvaraju, M., Sjoerds, Z., Smith, S. G., Smits, T., Spies, J. R., Sreekumar, V., Steltenpoh, C. N., Stenhouse, N., Swiatkowski, W., Vadillo, M. A., Van Assen, M. A. L. M., Williams, M. N., Williams, S. E., Williams, D. R., Yarkon, T., Ziano, I., and Zwaan, R. A. (2018). Justify your alpha. Unpublished manuscript, `https://psyarxiv.com/9s3y6/`.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day Inc., San Francisco, Calif.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation.* Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Lindley, D. V. (2014). *Understanding Uncertainty.* Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, revised edition.

Liu, C. and Martin, R. (2015). Frameworks for prior-free posterior probabilistic inference. *WIREs Comput. Stat.*, 7(1):77–85.

Martin, R. (2017a). A mathematical characterization of confidence as valid belief. Unpublished manuscript, `arXiv:1707.00486`.

Martin, R. (2017b). A statistical inference course based on *p*-values. *Amer. Statist.*, 71(2):128–136.

Martin, R. (2018). On an inferential model construction using generalized associations. *J. Statist. Plann. Inference*, 195:105–115.

Martin, R., Ermini Leaf, D., and Liu, C. (2012). Optimal inferential models for a Poisson mean. Unpublished manuscript, `arXiv:1207.0105`.

Martin, R. and Lin, Y. (2016). Exact prior-free probabilistic inference in a class of non-regular models. *Stat*, 5:312–321.

Martin, R. and Lingham, R. T. (2016). Prior-free probabilistic prediction of future observations. *Technometrics*, 58(2):225–235.

Martin, R. and Liu, C. (2013). Inferential models: a framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.*, 108(501):301–313.

Martin, R. and Liu, C. (2014). A note on p-values interpreted as plausibilities. *Statist. Sinica*, 24(4):1703–1716.

Martin, R. and Liu, C. (2015a). Conditional inferential models: combining information for prior-free probabilistic inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 77(1):195–217.

Martin, R. and Liu, C. (2015b). Marginal inferential models: prior-free probabilistic inference on interest parameters. *J. Amer. Statist. Assoc.*, 110(512):1621–1631.

Martin, R. and Liu, C. (2016). *Inferential Models*, volume 147 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL.

Martin, R. and Walker, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat.*, 8(2):2188–2206.

Martin, R. and Walker, S. G. (2017). Empirical priors for target posterior concentration rates. Unpublished manuscript, `arXiv:1604.05734`.

Mayo, D. (2014). On the Birnbaum argument for the strong likelihood principle. *Statist. Sci.*, 29(2):227–239.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge.* University of Chicago Press, Chicago.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing.* Cambridge University Press, Cambridge.

Mayo, D. G. and Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. In *Optimality*, volume 49 of *IMS Lecture Notes Monogr. Ser.*, pages 77–97. Inst. Math. Statist., Beachwood, OH.

McCullagh, P. (2002). What is a statistical model? *Ann. Statist.*, 30(5):1225–1310. With comments and a rejoinder by the author.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models.* Monographs on Statistics and Applied Probability. Chapman & Hall, London.

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. (2018). Abandon statistical significance. Unpublished manuscript, `arXiv:1709.07588`.

Meng, X.-L. (2017). X-L Files: Bayesian, Fiducial and Frequentist: BFF4EVER. *IMS Bulletin*, 46(4):17.

Molchanov, I. (2005). *Theory of Random Sets.* Probability and Its Applications (New York). Springer-Verlag London Ltd., London.

Nguyen, H. T. (1978). On random sets and belief functions. *J. Math. Anal. Appl.*, 65(3):531–542.

Nguyen, H. T. (2006). *An Introduction to Random Sets*. Chapman & Hall/CRC, Boca Raton, FL.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, 506:150–152.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).

Polyanin, A. D., Zaitsev, V. F., and Moussiaux, A. (2002). *Handbook of First Order Partial Differential Equations*, volume 1 of *Differential and Integral Equations and Their Applications*. Taylor & Francis Ltd., London.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Hutchinson and Co., Ltd., London.

Qiu, Y., Zhang, L., and Liu, C. (2018). Exact and efficient inference for partial Bayes problems. *Electron. J. Stat.*, 12(2):4640–4668.

Reid, N. and Cox, D. R. (2015). On some principles of statistical inference. *Int. Stat. Rev.*, 83(2):293–308.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, 12(4):1151–1172.

Savage, L. J. (1961). The foundations of statistics reconsidered. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 575–586. Univ. California Press, Berkeley, Calif.

Savage, L. J. (1972). *The Foundations of Statistics*. Dover Publications, Inc., New York, revised edition.

Savage, L. J. (1976). On rereading R. A. Fisher. *Ann. Statist.*, 4(3):441–500. Edited posthumously by John W. Pratt, With a discussion by B. Efron, Churchill Eisenhart, Bruno de Finetti, D. A. S. Fraser, V. P. Godambe, I. J. Good, O. Kempthorne and Stephen M. Stigler.

Scheffé, H. (1970). Practical solutions of the Behrens–Fisher problem. *J. Amer. Statist. Assoc.*, 65:1501–1508.

Schervish, M. J. (1996). *P* values: what they are and what they are not. *Amer. Statist.*, 50(3):203–206.

Schweder, T. and Hjort, N. L. (2002). Confidence and likelihood. *Scand. J. Statist.*, 29(2):309–332.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability*, volume 41 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, New York.

Seidenfeld, T. (1992). R. A. Fisher's fiducial argument and Bayes' theorem. *Statist. Sci.*, 7(3):358–368.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J.

Shafer, G. (1979). Allocations of probability. *Ann. Probab.*, 7(5):827–839.

Shafer, G. (2007). From Cournot's principle to market efficiency. In Touffut, J.-P., editor, *Augustin Cournot: Modelling Economics*, pages 55–95. Edward Elgar.

Shafer, G. (2016). *A Mathematical Theory of Evidence* turns 40. *Internat. J. Approx. Reason.*, 79:7–25.

Shafer, G. and Vovk, V. (2006). The sources of Kolmogorov's *Grundbegriffe. Statist. Sci.*, 21(1):70–98.

Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.*, 30:877–880.

Syring, N. and Martin, R. (2016). Robust and rate optimal gibbs posterior inference on the boundary of a noisy image. Unpublished manuscript, `arXiv:1606.08400`.

Syring, N. and Martin, R. (2017a). Calibrating general posterior credible regions. *Biometrika*, to appear, `arXiv:1509.00922`.

Syring, N. and Martin, R. (2017b). Gibbs posterior inference on the minimum clinically important difference. *J. Statist. Plann. Inference*, 187:67–77.

Taleb, N. N. (2018). *Skin in the Game: Hidden Asymmetries in Daily Life*. New York: Random House.

Taraldsen, G. and Lindqvist, B. H. (2017). Fiducial on a string. Unpublished manuscript, `arXiv:1706.03805`.

Trafimow, D., Amrhein, V., Areshenkoff, C. N., Barrera-Causil, C., Beh, E. J., Bigic, Y., Bono, R., Bradley, M. T., Briggs, W. M., Cepeda-Freyre, H. A., Chaigneau, S. E., Ciocca, D. R., Carlos Correa, J., Cousineau, D., de Boer, M. R., Dhar, S. S., Dolgov, I., Gomez-Benito, J., Grendar, M., Grice, J., Guerrero-Gimenez, M. E., Gutierrez, A., Huedo-Medina, T. B., Jaffe, K., Janyan, A., Liuzza, M. T., Lombardo, R., Marks, M. Meinlschmidt, G., Nalborczyk, L., Nguyen, H. T., Ospina, R., Suarez, I., Tegethoff, M., Tejo, M., van de Schoot, R., Vankov, I., Velasco-Forero, S., Wang, T., Yamada, Y., Zoppino, F. C., and Marmelojo-Ramos, F. (2018). Manipulating the alpha level cannot cure significance testing. *PeerJ Preprints*, 6:e3411v2, `https://doi.org/10.7287/peerj.preprints.3411v2`.

Trafimow, D. and Marks, M. (2015). Editorial. *Basic Appl. Soc. Psych.*, 37(1):1–2.

Valentine, J. C., Aloe, A. M., and Lau, T. S. (2015). Life after NHST: How to describe your data without "*p*-ing" everywhere. *Basic Appl. Soc. Psych.*, 37(5):260–273.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, volume 42 of *Monographs on Statistics and Applied Probability*. Chapman & Hall Ltd., London.

Walley, P. (2000). Towards a unified theory of imprecise probability. *Internat. J. Approx. Reason.*, 24(2-3):125–148. Reasoning with imprecise probabilities (Ghent, 1999).

Wasserman, L. A. (1990). Prior envelopes based on belief functions. *Ann. Statist.*, 18(1):454–464.

Wasserman, L. A. and Kadane, J. B. (1990). Bayes' theorem for Choquet capacities. *Ann. Statist.*, 18(3):1328–1339.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on *p*-values: context, process, and purpose [Editorial]. *Amer. Statist.*, 70(2):129–133.

Weerahandi, S. (1993). Generalized confidence intervals. *J. Amer. Statist. Assoc.*, 88(423):899–905.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29:350–362.

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34:28–35.

Woolston, C. (2015). Psychology journal bans *p* values. *Nature*, 519(9).

Xie, M.-g. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *Int. Stat. Rev.*, 81(1):3–39.

Yager, R. and Liu, L., editors (2008). *Classic Works of the Dempster–Shafer Theory of Belief Functions*, volume 219. Springer, Berlin.

Zabell, S. (1989). R. A. Fisher on the history of inverse probability. *Statist. Sci.*, 4(3):247–263. With comments by Robin L. Plackett and G. A. Barnard and a rejoinder by the author.

Zabell, S. L. (1992). R. A. Fisher and the fiducial argument. *Statist. Sci.*, 7(3):369–387.